TREBALL DE FI DE MÀSTER

Màster en Estadística i Investigació Operativa

Títol: AVALUACIÓ DELS ERRORS MOSTRALS DE L'ENQUESTA DE POBLACIÓ ACTIVA A CATALUNYA

Autor: SERGI MARTÍNEZ I MALDONADO

Directors: NÚRIA BOVÉ I FERRÉ, ENRIC RIPOLL I FONT

Empresa: INSTITUT D'ESTADÍSTICA DE CATALUNYA

Convocatòria: MAIG 2015

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Universitat Politècnica de Catalunya
Facultat de Matemàtiques i Estadística

Tesi de màster

Avaluació dels errors mostrals de l'Enquesta de Població Activa a Catalunya

Sergi Martínez i Maldonado

Directors: Núria Bové i Ferré, Enric Ripoll i Font

Institut d'Estadística de Catalunya

La veritable saviesa està en reconèixer la pròpia ignorància.

Sòcrates

Agraïments

A la Núria Bové, per tot el que m'ha aportat, tant a nivell acadèmic com a nivell personal.

A l'Enric Ripoll, per haver-me endinsat al món de l'estadística oficial, conjuntament amb la Manuela Alcañiz, i per les bones propostes que m'ha fet per millorar el treball.

A la Mònica Gasulla per la seva passió per la recerca i la generositat d'haver-me dedicat el seu temps.

A l'Idescat, que m'ha ofert la possibilitat de moure'm en un entorn professional càlid però rigorós, a banda d'oferir-me assistir a seminaris interessants. Als seus treballadors que m'han tractat molt bé i han fet que l'experiència fos agradable alhora que desafiant.

A la treballadora que hi havia al meu costat, companya de classe i bona amiga: la Jenny Montoya, qui ha cuidat de la meva integritat psicològica, i també física, oferint-me dolços i portant-me a restaurants llatinoamericans.

A l'Ana Alejandra, professora a qui aprecio i que m'ha ajudat amb els conceptes de remostreig, matèria que imparteix a la UAB.

A Carlos Pérez, por esas dudas resueltas por teléfono, y al INE en general.

To the Central Statistical Office of Poland for helping me.

Agraeixo als meus amics i familiars la comprensió del meu aïllament durant aquests mesos -gairebé un any- i al lector, per dedicar-li un temps a aquest treball que per mi ha estat tot una aventura.

Resum

Paraules clau: Teoria del mostreig, enquestes complexes, mètodes de remosteig, estadística oficial, Enquesta de població activa, força de treball

MSC2000: 62D05

Aquest treball s'emmarca dins de l'estadística oficial catalana i té com a objectiu final quantificar els errors de mostreig dels principals col·lectius poblacionals de l'Enquesta de població activa (EPA). Com és sabut, l'EPA és una operació per mostreig de caràcter continu realitzada per l'INE i ofereix dades per tot l'estat espanyol amb periodicitat trimestral. L'Idescat realitza trimestralment una important ampliació de resultats, tant en la desagregació de les categories d'algunes variables com en la incorporació de noves variables no tabulades per l'INE a nivell de Comunitats Autònomes. És en aquesta ampliació de resultats on sorgeix la necessitat de quantificar la precisió dels valors publicats per l'Idescat, és a dir, de quantificar els errors de mostreig associats a cada un dels col·lectius poblacionals i desagregant tant per les principals variables sociodemogràfiques (sexe, edat, nacionalitat, etc.) com de l'ocupació (sector d'activitat...).

Per tal de complir l'objectiu de l'encàrrec, s'ha fet un estudi dels mètodes existents i dels utilitzats per l'INE i altres Instituts d'estadística oficials en les estadístiques del mercat de treball. Amb l'objectiu d'assolir un aprenentatge òptim de cada un dels mètodes i de no descartar-ne cap dels principals a priori, s'han programat tots i s'han extret resultats per poder-los comparar, i així triar el mètode que ens sigui més adequat amb més coneixement de causa. No cal dir que estem treballant amb una enquesta complexa, amb un disseny mostral bietàpic, estratificat, amb resultats calibrats per diferents variables i on les especificacions en cadascuna de les fases de vegades ens poden portar dies de programació o d'indagacions vàries.

Superats aquests entrebancs, finalment hem aconseguit obtenir una bateria d'errors de mostreig per a cada col·lectiu i cadascun dels mètodes, i coincidint amb altres instituts d'estadística internacionals, triaríem la linealització de Taylor, un mètode que ha resultat ser estable al llarg del temps i que ve recolzat per les propietats del mètode de Jackknife.

Així doncs, a banda d'obtenir les estimacions per diferents magnituds del mercat laboral, i segons diverses variables sociodemogràfiques o de l'ocupació, podríem aplicar aquesta metodologia en altres enquestes. També aconseguim una forma de poder facilitar un càlcul bastant precís dels errors a un possible investigador, sense violar el secret estadístic.

Abstract

Keywords: Sampling theory, complex surveys, resampling methods, official statistics, labour force survey, workforce

MSC2000: 62D05

This essay is framed within Catalan official statistics and aims to finally quantify the ammount of sampling errors on main population groups in the Labour Force Survey (In Catalan: EPA). As is generally known, EPA is a continuous sampling survey carried out by INE and provides with quarterly data from all around Spain. Idescat enhances these results quarterly, both by disaggregating categories from some of the variables and by adding new variables about autonomous communities that are not tabulated by INE. This enhancement of results arises the need of quantifying Idescat's published values accuracy; in other words, quantifying sampling errors that are related to each population group and categorising both by sociodemogrpahic variables (sex, age, nationality, etc.) and by occupation (activity sector...).

In order to accomplish the goals of this task, several existing methods on labour market surveys, including those used by INE and other official statistical institutes, have been studied. So as to achieve an optimal learning of each method and no to discard, a priori, any of the main ones, all methods have been programmed and results have been extracted so they could be compared. As a result, the most appropriate one may be chosen with full information. It goes without saying that we are dealing with a complex survey featuring a two-stage stratified sample design. The results have been calibrated using different variables. Besides, the specifications of each phase could sometimes involve several days programming and investigating.

Once these obstacles have been overcome, an amount of sampling errors have been finally obtained for each population group and method. Taylor's linearisation has been chosen, in coincidence with other international statistical institutes, as it is a method that has proven to be stable over time and is supported by Jackknife's method properties.

Thus, besides obtaining estimations for different labour market aspects and by different sociodemographic or occupational variables, this methodology could be used in other surveys. In addition, a way to make a quite precise error calculus easier for any other investigator is achieved without violating statistical secrecy.

Abreviatures

BRR	*Balanced Repeated Replication* (mostres equilibrades)
CV	Coeficient de variació
DEFF	Efecte del disseny
E()	Esperança matemàtica
EPA	Enquesta de Població Activa
EQM	Error quadràtic mitjà
H-T	Horvitz-Thompson
IC	Interval de confiança
Idescat	Institut d'Estadística de Catalunya
INE	*Instituto Nacional de Estadística*
m.a.s.	Mostra aleatòria simple
min	Minimitzar (problema d'optimització)
OIT	Organització Internacional del Treball
PSU	*Primary Sampling Units* (unitats mostrals primàries)
s.a.	Subjecte a (problema d'optimització)
SE	*Standard Error* (error estàndard)
SSU	*Secondary Sampling Units* (unitats mostrals secundàries)
Var()	Variància

Índex general

Índex de figures

Índex de taules

Presentació

L'objectiu del present treball és discutir diferents mètodes per estimar l'error mostral d'una enquesta complexa com ho és la EPA i facilitar els resultats aplicats a un trimestre. Per facilitar-ne la lectura, l'hem dividit en tres parts. Els lectors que no estiguin familiaritzats amb les enquestes de força de treball, amb els conceptes estadístics de mostreig i de remostreig, o amb la problemàtica en que se centra el treball, poden començar per la primera part. La segona part explica al lector el disseny i metodologia actual i la tercera és la proposta que hem estudiat per al cas de l'Enquesta de població activa a Catalunya.

La **Part 1** és el marc teòric i la contextualització, per tal de situar-nos dins el problema plantejat, i comprèn els dos primers capítols. Al **Capítol 1**, s'expliquen els objectius de l'EPA i la seva importància, tot situant-la en el context de l'estadística oficial i comparant-la amb les demés fonts estadístiques de la força de treball. El **Capítol 2** és de caire teòric i introdueix al lector en les tècniques de mostreig i de remostreig, essencials per avançar en la lectura.

La **Part 2** presenta el disseny de l'EPA i les principals mesures d'interès, i conté els capítols 3 i 4. El **Capítol 3** és una explicació exhaustiva de com està dissenyada aquesta enquesta. El **Capítol 4** mostra com l'INE i altres oficines d'estadística oficial obtenen els resultats i la seva fiabilitat.

La **Part 3** és la més extensa, comprenent els cinc darrers capítols, i és on presentem la nostra proposta metodològica. Al **Capítol 5**, descrivim les bases de dades que fem servir per a l'estudi. Al **Capítol 6**, mostrem la metodologia per tal d'estimar les magnituds de treball i el seu error, que serà discutida al **Capítol 7**. Els resultats es recullen al **Capítol 8**. Finalment, el **Capítol 9** és la conclusió del treball, on avaluem què hem aportat i on podríem seguir investigant.

Desitgem que sigui una lectura agradable i del seu interès.

Part 1

Introducció i marc teòric

Capítol 1
L'Enquesta de població activa

En el món en què vivim actualment, no es poden prendre decisions encertades ni es pot conèixer la realitat d'un país o territori, si les nostres afirmacions no estan fonamentades en dades fiables. Una de les necessitats d'informació que a hores d'ara ningú qüestiona, és la de conèixer les múltiples situacions que poden tenir els individus en relació amb l'activitat econòmica i aquest coneixement requereix una operació feta a mida amb un qüestionari específic i detallat. En general és difícil que aquest objectiu sigui assolit amb èxit per una operació que tingui una finalitat més general o per l'explotació d'un registre administratiu, i és per aquest motiu que la majoria de països, seguint les recomanacions de l'Organització Internacional del Treball (OIT), disposen d'una operació estadística específica destinada a cobrir les necessitats d'informació en aquest àmbit.

1. Marc legal

En el cas de la Unió Europea, els països membres estan obligats a complir el Reglament(CE) núm 577/98 del Consell de 9 de març de 1998 relatiu a l'organització d'una enquesta mostral sobre la població activa a la Comunitat. És a dir, estan obligats a dur a terme una enquesta sobre la força de treball d'acord amb un qüestionari mínim i una metodologia fixats per Eurostat. En el cas espanyol, i per tant de Catalunya, les dades de l'enquesta comunitària sobre la força de treball es basen en els resultats de l'Enquesta de població activa (EPA) que elabora l'INE trimestralment per a tot el territori.
L'Enquesta de població activa (EPA) és la principal font d'informació estadística del mercat de treball. L'EPA és una investigació per mostreig dirigida a la població resident en habitatges familiars de caràcter permanent o habitual, contínua i amb l'objectiu principal de conèixer la relació amb l'activitat econòmica de la població. És a dir, ofereix resultats dels principals col·lectius poblacionals en relació amb el mercat de treball (actius, ocupats, assalariats, desocupats i inactius) classificats segons les principals característiques sociodemogràfiques i de l'ocupació de la població.
Les definicions que aplica l'EPA es basen en les diferents recomanacions de l'OIT i els criteris utilitzats per obtenir la informació són coherents amb els establerts pels

organismes internacionals, fet que permet que els resultats siguin comparables amb els d'altres països i amb el conjunt d'Europa. A través de l'explotació estadística dels fitxers de microdades per a Catalunya de l'Enquesta de població activa que l'INE envia trimestralment a l'Idescat, s'assoleix l'actuació Estadística de població activa inclosa en els diferents plans anuals d'actuació estadística que desplega l'activitat Estadística de força de treball prevista en la Llei 13/2010, de 21 de maig, del Pla estadístic de Catalunya 2011–2014. La Llei 23/1998 d'estadística de Catalunya, de 30 de desembre, estableix que les estadístiques oficials han de complir un seguit de requisits:

- Ser declarades d'interès públic pel Parlament de Catalunya, mitjançant la inclusió del Pla estadístic de Catalunya. Aquest Pla preveu les activitats estadístiques que s'han de dur a terme al llarg de quatre anys i es desenvolupa amb els programes anuals d'actuació estadística.
- Ser objectives en el seu plantejament i en els seus resultats.
- Ser comparables amb altres estadístiques d'àmbit estatal o europeu, aplicant un sistema normalitzat de conceptes, definicions, classificacions i codis, així com una metodologia que permeti aquesta comparabilitat.
- Garantir la no-duplicitat amb altres estadístiques existents per tal d'evitar molèsties innecessàries als ciutadans i una millor distribució dels recursos públics.
- Respectar la intimitat personal, assegurar el secret estadístic i permetre als interessats conèixer les característiques de l'enquesta.

És per tots aquests requisits que estableix la citada llei, que l'Idescat realitza una ampliació de resultats de l'EPA, amb la informació base provinent de l'INE, utilitzant les definicions recomanades internacionalment a l'hora de calcular els indicadors i els col·lectius poblacionals i aplica criteris per respectar el secret i la qualitat estadística dels resultats.

El concepte de qualitat estadística el trobem recollit a la normativa europea, tal com s'expressa en l'article 12 del Reglament (CE) No 223/2009 del Parlament Europeu i del Consell d'11 de març de 2009, relatiu a l'estadística europea per tal de garantir la qualitat dels resultats de les estadístiques europees. Aquestes estadístiques es desenvoluparan, elaboraran i difondran segons normes uniformes i mètodes harmonitzats. El mateix reglament ens especifica els criteris de qualitat a seguir:

a) Pertinència: grau en que les estadístiques responen a necessitats actuals i potencials dels usuaris. Com ja s'ha explicat anteriorment, ningú dubta de la necessitat de tenir dades fiables de la relació amb l'activitat econòmica de la població i per tant de l'elaboració d'una Enquesta de població activa.
b) Precisió: concordança de les estimacions amb els valors reals desconeguts. És en aquest punt on el treball que aquí es presenta té una especial importància, ja que un dels seus objectius és "quantificar" la precisió dels resultats de l'Enquesta de població activa obtinguts per a Catalunya.
c) Puntualitat i actualitat: temps transcorregut entre la data de publicació dels resultats i la data de referència dels mateixos. Aquest és un dels punts forts de l'Enquesta de població activa, ja que les dades del trimestre t es publiquen a t+1, és a dir, només un mes després.

d) Accessibilitat i claredat: condicions i modalitats en que els usuaris poden obtenir, utilitzar i interpretar els resultats. L'Idescat posa a disposició dels usuaris una base de dades on consultar l'ampliació de resultats que s'elabora trimestralment, així com tota una sèrie d'indicadors amb l'objectiu de facilitar la comparativa amb dades d'Espanya i de la Unió Europea.

   - Vincle base de dades EPA:
     http://www.idescat.cat/treball/epa

   - Vincle indicadors conjuntura econòmica:
     http://www.idescat.cat/economia/inec?tc=2& id=06

   - Vincle indicadors estructura econòmica:
     http://www.idescat.cat/economia/inec?tc=2& id=57

   - Vincle indicadors europeus:
     http://www.idescat.cat/economia/inec?tc=2& id=82

e) Comparabilitat: mesura de l'impacte de les diferències en els conceptes aplicats i en els instruments i procediments de mesura quan es comparen estadístiques entre diferents zones geogràfiques, àmbits sectorials o al llarg del temps. L'Idescat quan realitza actualitzacions de sèries a causa de la revisió de les estimacions poblacions o algun canvi de definició d'alguna variable rellevant, sempre intenta posar a l'abast de l'usuari una mesura per poder quantificar el canvi en qüestió.

Així doncs, tal com mostra el citat article del Reglament (sobretot l'apartat b), com el principi número 12 (precisió i fiabilitat) del codi de bones pràctiques europees, aquest treball té una importància remarcable en el procés de difusió dels resultats de l'EPA, ja que pot ajudar a redefinir el pla de tabulació en el cas que els mètodes mostrin un errors de mostreig no desitjables en algunes taules, alhora que l'estudi d'aquests errors aportarà informació rellevant en el moment d'interpretar els resultats de l'EPA.

2. Definicions dels principals col·lectius poblacionals de l'EPA

Els col·lectius principals que hem de tenir en compte per analitzar la força de treball són quatre: població activa, població ocupada, població desocupada i població inactiva.

La **població activa** la podem definir com el conjunt de persones que subministren mà d'obra per a la producció de béns i serveis o bé estan disponibles i fan gestions per incorporar-se a la producció. És a dir, la població activa està formada per la població ocupada més la desocupada.

L'indicador relatiu associat a la població activa és la **taxa d'activitat**, que es

defineix de la següent forma:

$$taxa\ d'activitat = \frac{població\ activa}{població\ de\ 16\ anys\ i\ més} \cdot 100$$

La **població ocupada** és la població de 16 anys o més que ha treballat, per compte d'atri o per compte propi, a la seva ocupació principal o a una de secundària. Per "treballar" s'ha d'entendre sempre com la realització d'una activitat a canvi d'un sou, salari, benefici empresarial o guany familiar, en metàl·lic o en espècie, durant una setmana concreta (l'anterior a la de l'entrevista), durant almenys una hora. Quan un individu no compleix aquestes condicions, es diu que "no ha treballat" o que està "absent de l'ocupació".
L'indicador relatiu associat a la població ocupada és la **taxa d'ocupació**, que es defineix de la forma següent:

$$taxa\ d'ocupació = \frac{població\ ocupada}{població\ de\ 16\ anys\ i\ més} \cdot 100$$

Un subcol·lectiu de la població ocupada que sempre mereix una atenció especial és la **població ocupada assalariada**. Aquest subcol·lectiu s'obté classificant la població segons la situació professional.
L'indicador relatiu associat a la població ocupada assalariada és la **taxa de salarització**, que es defineix de la forma següent:

$$taxa\ de\ salarització = \frac{població\ ocupada\ assalariada}{població\ ocupada} \cdot 100$$

La **població desocupada** és la població de 16 anys o més que no ha treballat, està disponible per treballar i busca una ocupació. És important notar que s'han de complir les tres condicions per poder classificar una persona com a desocupada. Es consideren disponibles les persones que podrien començar a treballar en el termini de dues setmanes (a partir de la data de l'entrevista). Es considera que hi ha una cerca efectiva d'ocupació quan s'han efectuat gestions en aquest sentit o per tal d'establir-se per compte propi durant les quatre setmanes anteriors (a la data de l'entrevista).
L'indicador relatiu a la població desocupada és la **taxa d'atur**, que es defineix de la forma següent:

$$taxa\ d'atur = \frac{població\ desocupada}{població\ activa} \cdot 100$$

La **població inactiva** està formada per aquelles persones que no treballen i que o bé no cerquen ocupació o bé no estan disponibles per treballar.

3. L'EPA versus altres fonts de dades del mercat de treball

A banda de l'EPA, que és l'enquesta dissenyada específicament per estudiar la relació amb l'activitat de la població, hi ha altres operacions estadístiques sobre les famílies i les empreses que també aporten informació molt rellevant.

Els censos de població i habitatges són operacions que inclouen la relació amb l'activitat com una variable més de classificació. Faciliten l'avaluació dels principals col·lectius i les taxes associades, i una de les característiques més importants és que permeten fer-ho amb una alta desagregació territorial. Fins el Cens de població i habitatges del 2001, la desagregació territorial era màxima. A partir del Cens del 2011, hi ha certes restriccions derivades de la introducció d'una part mostral a l'hora d'obtenir les dades.

A banda dels Censos i de les enquestes, també s'obtenen dades del mercat de treball de l'aprofitament de registres administratius, que permeten obtenir informació puntual, molt exhaustiva i amb un cost reduït. Ara bé, cal tenir en compte que els registres administratius estan pensats per a la gestió administrativa i no per a la producció d'informació estadística, i és per aquest motiu que les dades són molt sensibles als canvis de gestió i normativa, i la comparabilitat a escala internacional resulta molt difícil.

Així doncs, com a alternativa als resultats de l'atur estimat que ens ofereix l'EPA, és habitual trobar dades d'aquest col·lectiu extretes dels aturats inscrits a les oficines públiques d'ocupació o del Cens de població i habitatges.

Ara bé, cal tenir en compte que les tres fonts de dades ofereixen resultats diferents, ja que en el fons, també quantifiquen col·lectius diferents.

Atur estimat EPA: Tal com ja s'ha definit més amunt, per poder classificar una persona com a desocupada, ha de complir tres condicions: no haver treball, estar disponible per treballar i realitzar una cerca activa de feina.

Atur Registrat: població que no està treballant i que està inscrita com a demandant d'ocupació a les oficines públiques d'ocupació. Ara bé, és important saber que s'exclouen col·lectius gens insignificants. Alguns dels més rellevants són:

- Aquells que busquen una ocupació de poques hores (<20 hores setmanals)
- Estudiant menors de 25 anys
- Estudiants majors de 25 anys i que busquen la primera feina
- Aquells que busquen una ocupació de caràcter conjuntural (<3 mesos)

Atur provinent del Cens de població i habitatges: respon al resultat de l'explotació de la pregunta mostrada a la FIGURA 1.1.

Però encara que pugui semblar que els resultats haurien de ser semblants als de l'EPA, no ho són perquè aquí l'entrevistat s'autoclassifica i, en canvi, a l'EPA la relació amb l'activitat es dedueix a partir de les respostes d'una àmplia bateria de preguntes que determinen si la cerca d'ocupació és activa i la disponibilitat real d'incorporar-se al mercat de treball de l'entrevistat.

Pel que fa a les dades de la població ocupada, l'explotació estadística del registre d'afiliats en alta laboral a qualsevol dels règims de la Seguretat Social constitueix la principal font d'origen administratiu per estimar l'ocupació. Cal advertir que, com passa en les enquestes a les empreses, el concepte estimat no és el de treballadors, sinó el de llocs de treball i que els resultats no coincideixen amb els de l'EPA perquè aquí, com és lògic, no es recull el que es coneix com a economia submergida.

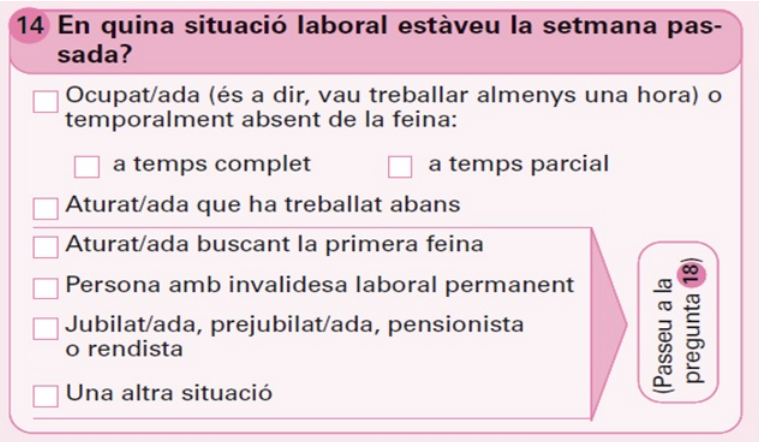

FIGURA 1.1. Model de pregunta del Cens.

Capítol 2
Marc teòric

Si volem obtenir una informació d'una població, sovint hem de recórrer a agafar-ne un subconjunt d'aquesta per tal de fer el problema assumible. Aquest subconjunt s'anomena **mostra** i és rellevant saber com triar-lo per tal de trobar la dada de la forma més acurada possible i sense deixar de banda el cost de la seva obtenció.

1. Tipus de mostreig

El mostreig és el procediment amb el qual se selecciona una mostra. Segons si aquesta tria és sotmesa a criteris probabilístics o no, el classificarem en **mostreig probabilístic** i en **mostreig no probabilístic**, respectivament.
Per a les enquestes oficials, s'adopten els primers, dels quals n'hi ha de diversos tipus, com ara:

- **Mostreig aleatori simple**: D'una població finita d'N unitats, se n'extreu una quantitat anteriorment fixada i de tal forma que totes les unitats tenen la mateixa probabilitat de ser triades. Normalment se sobreentén que es fa **sense reemplaçament**, és a dir, que els individus poden ser triats com a màxim una vegada.
- **Mostreig sistemàtic**: Les N unitats estan ordenades i se n'extrauran n (les lletres majúscules faran referència a les dades poblacionals i les minúscules, a les mostrals). La primera, amb posició m, és triada a l'atzar i les demés es trien agafant l'element a la posició $m + k$, on $k = N/n$, i un cop s'assoleix el final de la llista es torna a l'inici un cop, sense arribar a la posició m de nou. D'aquesta forma les unitats són equidistants, s'incrementa la precisió i es procura aconseguir dades el màxim d'heterogènies. Sovint s'utilitza també en cas de distribucions no equiprobables, replicant les unitats un nombre de cops proporcional al seu pes.
- **Mostreig per conglomerats**: Si la població està dividida en grups que contenen tota la variabilitat d'aquesta, es pot plantejar fer la tria d'alguns d'aquests grups. L'estimació pot ser més precisa si la major part de la variabilitat és dins del grup, i no entre els grups.
- **Mostreig estratificat**: Per quan es pot classificar la població en diverses subpoblacions (estrats) que comparteixin alguna característica de manera que la variabilitat dins del grup sigui mínima i entre els estrats el més gran possible.

Això pot millorar la precisió de les estimacions a nivell poblacional i permetre estimar a nivell subpoblacional.

En el cas del mostreig no probabilístic n'hi ha, entre d'altres:

- **Mostreig per quotes**: És semblant al mostreig estratificat. La mostra se selecciona en un nombre proporcional a aquells que compleixen una característica d'una població (normalment edat i sexe).
- **Mostreig opinàtic**: Basat en l'opinió d'un expert.
- **Mostreig de bola de neu**: Els propis entrevistats conduiran a individus que s'inclouran. És indicat per a poblacions minoritàries, clandestines o molt disperses però en contacte entre si.

Com s'ha citat anteriorment, el procés es pot fer **amb reemplaçament** o sense, depenent de si un individu ja agafat a la mostra pot tornar a ésser elegit, o no.
Una altra distinció possible es fa entre si el mostreig és **directe** o **indirecte**. En el primer cas, surten elegides les unitats mostrals més elementals, mentre que en el segon cas, a partir de les seleccionades, s'agafen totes les del conjunt on aquestes pertanyen.
El procés de mostreig pot esdevenir més complex si es combinen diversos conceptes definits anteriorment:
Un **mostreig en etapes múltiples** consisteix en obtenir una mostra d'unitats primàries de mostreig i de cadascuna obtenir, mitjançant un nou mostreig, les unitats secundàries de mostreig, i així successivament fins arribar a les unitats més elementals.
Un **mostreig multifàsic** es basa en que certs elements d'informació procedeixen de tota la mostra, però d'altres es prenen de submostres. Per exemple, en una primera fase pot interessar identificar tots aquells individus que han comprat un tipus de producte i, en una fase posterior, conèixer informació addicional sobre aquest subgrup, com podria ser la marca del producte que han obtingut.
Direm que els dos anteriors són **dissenys mostrals complexos**.

2. Conceptes de disseny mostral

En una enquesta, l'**àmbit** abasta tres òptiques diferents: **el poblacional** (població objecte de l'estudi), el **geogràfic** (territori segons l'objectiu de l'enquesta) i el **temporal** (període de referència de l'enquesta i període de referència de la informació o presa de dades).
Segons la permanència dels elements mostrals de l'enquesta, es poden classificar en:

- Enquestes **contínues de panel rotant**: Proporciona informació periòdica de manera que facilita estimacions transversals d'un moment temporal concret i a més, permeten d'estudiar l'evolució de les característiques poblacionals a estudiar. Entre una enquesta i la següent, romana una part comuna que com més gran és, més precisió donen a l'estimació dels canvis. Una part, doncs, és renovada i així s'evita enquestar les mateixes persones perpètuament.
- Enquestes **contínues de panel fix**: En aquest cas la mostra és sempre la mateixa. Malgrat que pot semblar ideal per estimar els canvis, pot resultar complicat realitzar el seguiment de totes les unitats mostrals i calen fer ajustos probabilístics en els casos d'incorporacions o baixes d'aquestes al panel.

- Enquestes **esporàdiques**: Tenen la finalitat de donar una informació transversal de les variables objecte d'estudi referida a un temps concret. Com no té continuïtat, permet una gran flexibilitat en el disseny.

El **marc** és l'eina que s'utilitza per tenir accés a la població i per tant, poder seleccionar la mostra. N'hi ha de dos tipus:

- De llista: dóna accés directe als individus (e.g., un llistat dels estudiants d'una universitat).
- De zona: dóna una llista de les àrees geogràfiques on hi ha accés als individus de forma indirecta (e.g., els barris de la ciutat de Barcelona -els noms individuals s'hauran d'esbrinar-).

A més a més, pot incloure la **informació complementària** que pot emprar-se amb la finalitat de millorar el disseny mostral com, per exemple, en el procés d'estimació, el d'estratificació o en l'ajust de la manca de resposta.
Es defineix l'**afixació** com la distribució de la mostra en funció dels diferents estrats. Pot ser de diversos tipus:

- Uniforme o simple: Tots els estrats tindran la mateixa grandària.
- Proporcional: La grandària mostral de l'estrat s'assigna en funció al pes que representa aquell estrat sobre el total poblacional. És la més intuïtiva i utilitzada, però pot deixar de banda aquells sectors on el pes a la població és baix.
- De mínima variància: És difícil d'aplicar i es basa en agafar menys mostra dels estrats que siguin més semblants, i a l'inrevés.

El **factor d'elevació** és el nombre d'elements de la població que és representat per cada element de la mostra. La suma dels factors d'elevació és igual al total de la població (N). El **pes** és el factor d'elevació ajustat de tal forma que la suma sigui igual als elements de la mostra (n). És necessari treballar amb pesos quan es fan inferències, per tal de no infraestimar els errors estàndard. El mètode de **ponderació** serveix per ajustar els pesos d'una mostra utilitzant informació auxiliar. D'aquesta manera, l'enquesta reflectirà bé el pes que cada grup (per exemple, d'edat) té en la població. Per abús del llenguatge, sovint utilitzarem com a equivalents els mots factors, pesos i ponderacions, fent referència als factors d'elevació.

3. Expressió de l'error

Quan es dóna una estadística, i més en l'entorn de l'estadística oficial, la precisió és un requisit imprescindible. D'aquí rau la importància de conèixer l'error que es comet en presentar la dada, essent també d'interés la possibilitat de controlar-lo o minimitzar-lo, i no només d'acotar-lo.
En aquesta secció expliquem les mesures d'error més comunes. A continuació es presenten d'una forma molt senzilla alguns conceptes bàsics per poder interpretar les mesures de l'error de la mostra.
Un **estimador** $\hat{\theta}$ és una funció de la mostra creat per esbrinar un **paràmetre** desconegut θ de la població.
Hom pot definir l'**error mostral** com la imprecisió que es comet en estimar una característica de la població de l'estudi mitjançant el valor que s'ha obtingut a partir d'una part o mostra de la població. Tractant-se de la dispersió sobre un

estimador, es dóna l'expressió generalitzada de l'error mostral:

$$\text{Error de mostreig} = \sqrt{Var(\hat{\theta})},$$

on $Var(\hat{\theta})$ fa referència a la **variància** de l'estimador, és a dir, $\text{Var}(\widehat{\theta}) = \text{E}[(\widehat{\theta} - \text{E}(\widehat{\theta}))^2]$.

$\text{E}(\widehat{\theta})$ és l'**esperança matemàtica** de l'estimador o valor esperat.

El **biaix** d'un estimador és la diferència entre la seva esperança matemàtica i el valor del paràmetre que estima:

$$\text{Biaix}[\,\hat{\theta}\,] = \text{E}[\,\hat{\theta}\,] - \theta = \text{E}[\,\hat{\theta} - \theta\,],$$

És desitjable que els estimadors siguin no-esbiaixats, és a dir, que tinguin un biaix igual a zero.

L' **Error Quadràtic Mitjà (EQM)** d'un estimador, és el valor esperat dels errors quadràtics, però es pot expressar també com a la suma següent:

$$\text{EQM}(\hat{\theta}) = \text{Var}(\hat{\theta}) + \Big(\text{Biaix}[\hat{\theta}]\Big)^2$$

Per tant, en cas de trobar-nos amb dos estimadors no esbiaixats, els EQM coincidirien amb la variància de cada estimador: el de menor EQM serà el de menor variància.

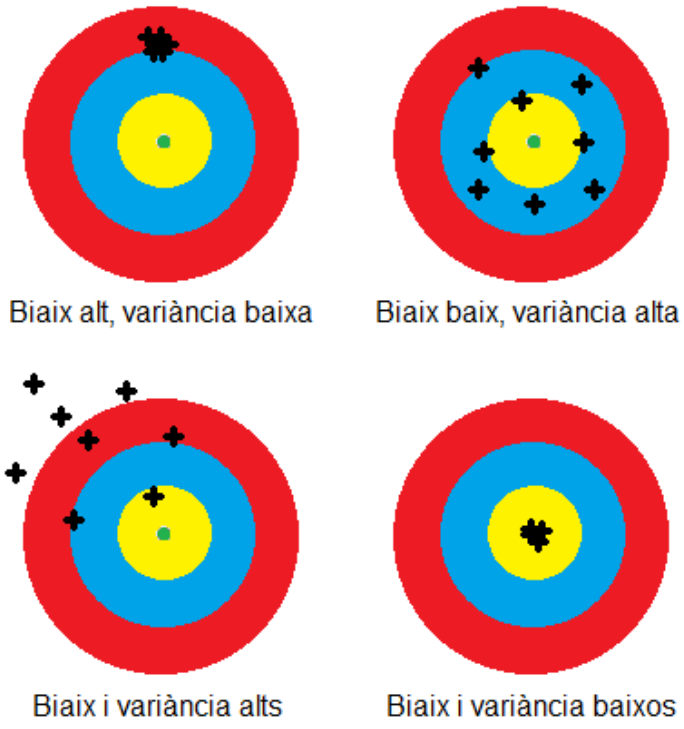

FIGURA 2.1. Biaix i variància

El **coeficient de variació** (**CV**) és una mesura de dispersió relativa de gran utilitat per a la comparació dels errors, problema d'abordament molt restringit amb les expressions mostrades anteriorment. El CV és expressat com a:

$$CV(\hat{\theta}) = \frac{\sqrt{Var(\hat{\theta})}}{\hat{\theta}} \cdot 100$$

si es vol mostrar en forma percentual.

L'**efecte del disseny** és la raó entre la variància de l'estimador basat en un disseny complex i la variància de l'estimador basat en una mostra aleatòria simple (m.a.s.) de la mateixa grandària. En el cas de les mostres estratificades, haurà de ser menor a 1: recordem que no pot ser menys eficient que una mostra aleatòria simple. En canvi, en el mostreig per conglomerats sí que serà igual o superior a 1.

$$DEFF(\hat{\theta}) = \frac{Var(\hat{\theta}_{disseny})}{Var(\hat{\theta}_{m.a.s.})}$$

Al capítol 4.3, es tractaran els factors que comporten l'error.

Per últim, comentem que els **intervals de confiança** són també molt utilitzats per dades que provenen de mostreigs complexos, i es basen en les distribucions Normal(0,1) i t de Student.

4. Tècniques per a l'avaluació de l'error mostral. El remostreig

En aquest treball s'ha optat per utilitzar tècniques de remostreig per tal d'estimar l'error mostral. Aquestes es basen en fer submostres de la mostra per construir distribucions de l'estadístic d'interès.

A continuació explicarem breument els mètodes clàssics de remostreig. Els primers (de grups aleatoris i de repeticions equilibrades) van sortir del treball de camp, de la necessitat de trobar en un moment donat una solució que era creativa i que semblava lògica. Posteriorment, es van trobar certes deficiències i els mètodes van ésser millorats. Els altres mètodes van sorgir en el marc de població infinita: són el Jackknife i el Bootstrap. El mètode de linealització també va néixer en aquest marc i, malgrat no ser un mètode de remostreig, també el tractarem.

4.1. Els grups aleatoris.

Poden ser dependents o independents. En ambdós casos, es fan particions aleatòries de la mostra en R grups. L'estimador $\hat{\theta}$ es forma d'aquesta manera:

$$\hat{\theta}_{(.)} = \frac{1}{R} \sum_{r=1}^{R} \hat{\theta}_{(r)},$$

on $\hat{\theta}_{\cdot}$ és la mitjana dels $\hat{\theta}_r$.
La variància de l'estimador la dóna l'expressió:

$$\hat{V} = \frac{1}{R(R-1)} \sum_{r=1}^{R} (\hat{\theta}_{(r)} - \hat{\theta}_{(.)})^2$$

4.1.1. *Pros.*

- Fàcils de calcular.
- Mètode general.

4.1.2. *Contres.*

- Si s'assumeixen grups independents, pot ser un problema quan N és petita.
- Si són grups dependents, l'estimador del paràmetre i de la variància seran esbiaixats.
- Pot no ser assumible si es desitja una bona mida mostral a cada grup.
- Obsolet. Els mètodes presentats a continuació estenen la idea del mètode dels grups aleatoris.

4.2. Semimostres reiterades i BRR.

És una versió prèvia i no equilibrada del *Balanced Repeated Replication* (BRR), que serà abordat a la proposta metodològica del Capítol 6.
La seva versió més simple particiona la mostra en dues submostres R cops, obtenint 2R semimostres. La variància s'estima com segueix:

$$\hat{V}(\hat{\theta}) = \frac{1}{2R} \sum_{i=1}^{2R} (\hat{\theta}_i - \hat{\theta})^2,$$

on $\hat{\theta}_i$ és l'estimació de cada semimostra i i $\hat{\theta}$ és la del total.

4.2.1. *Pros.*

- Es pot estendre per utilitzar pesos.
- Si s'equilibra, és asimptòticament equivalent a mètodes de linealització per funcions suaus de totals poblacionals i quantils.
- Si s'equilibra, el seu cost computacional serà relativament baix.
- Tracta la no-resposta de manera senzilla.

4.2.2. *Contres.*

- 2 PSU per mostra (pot estendre's amb mètodes més complexos)
- Si no és equilibrat, no compleix les propietats d'ortogonalitat que es comentaran a la proposta metodològica.
- Si no s'equilibra, pot tenir un cost computacional molt alt, depenent de les iteracions que es facin.

4.3. El Jackknife.

Es basa en crear n replicacions, tantes com unitats hi hagi, i en cadascuna n'eliminarem una. Es pot generalitzar, i considerar que les seves unitats a extreure són

les PSUs. La seva variància vindrà donada, doncs, per:

$$\hat{V}(\hat{\theta}) = \sum_h \frac{n_h - 1}{n_h} \sum_{j=1}^{n_h} (\hat{\theta}_{hj} - \hat{\theta})^2.$$

$\hat{\theta}_{hj}$ és l'estimació del total on s'elimina la PSU j de l'estrat h.
$\hat{\theta}$ és l'estimació del total.

4.3.1. *Pros.*

- Pot tractar bastants estimadors.
- Senzill d'implementar a qualsevol disseny.
- Tracta la no-resposta de manera fàcil.

4.3.2. *Contres.*

- Esbiaixat si hi ha poca variabilitat entre les diferents PSUs o si hi ha una gran variabilitat interna.

4.4. El Bootstrap.

La versió més senzilla consisteix en:

(1) Obtenir una m.a.s. amb reemplaçament de la mateixa mida que la mostra.
(2) Calcular l'estimador.
(3) Repetir la rutina M cops.
(4) L'estimació de la variància serà:

$$\hat{V}(\hat{\theta}) = \frac{1}{M-1} \sum_{k=1}^{M} (\hat{\theta}_k - \hat{\theta})^2.$$

4.4.1. *Pros.*

- Permet de tractar amb la majoria dels estimadors.
- Aconsegueix una aproximació de la distribució de l'estadístic d'interès.
- Robust. Bona alternativa als estimadors paramètrics.

4.4.2. *Contres.*

- S'han fet estudis on no és recomanable a gran escala.
- Més complicat d'aplicar que el Jackknife.
- Requereix d'un ordinador més potent.

4.5. La linealització de Taylor: Un altre enfocament.

Permet de calcular estimacions de l'error mostral per a totals, mitjanes i ràtios en mostres amb estratificació, conglomerats i probabilitats desiguals. El mètode obté aproximacions lineals de l'estimador i calcula la seva variància (i en conseqüència, l'error mostral).
N'hi ha de diversos tipus. Com a exemple, la variància de la mitjana poblacional amb una **expansió de Taylor** es calcula mitjançant l'expressió que segueix:

$$\hat{V}(\hat{Y}) = \sum_{h=1}^{H} \frac{n_h(1-f_h)}{n_h - 1} \sum_{i=1}^{n_h} (e_{hi.} - \bar{e}_{h..})^2,$$

on

$$e_{hi.} = \frac{\sum_{j=1}^{m_{hi}} w_{hij}(y_{hij} - \hat{Y})}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}}$$

i

$$\bar{e}_h = \frac{\sum_{j=1}^{n_h} e_{hi.}}{n_h}.$$

h són els estrats, i els conglomerats dins l'estrat, i j la unitat dins el conglomerat i de l'estrat h. $n = \sum_{h=1}^{H}\sum_{i=1}^{n_h} m_{hi}$ és el nombre total d'observacions a la mostra.

4.5.1. *Pros.*

- Extensa aplicabilitat
- Teoria ben desenvolupada
- Gran quantitat de *software* disponible

4.5.2. *Contres.*

- Depenent del mètode, pot requerir de trobar derivades parcials difícils d'obtenir.
- Per cada estadístic es necessita un mètode diferent.
- Tot i que és possible tractar la no-resposta, fer-ho és complicat.

Part 2

L'EPA: Disseny i metodologia

Capítol 3
El disseny mostral de l'EPA

1. Àmbit

1.1. Àmbit poblacional.

L'enquesta s'adreça a la població resident en habitatges familiars principals i allotjaments fixos. S'hi exclouen, per tant, els habitatges col·lectius (hospitals, hotels, convents, etc.), que representen al voltant d'un 1% de la població total. S'hi inclouen els estrangers que pensen residir a Espanya un mínim d'un any, aquells espanyols que han marxat menys d'un any fora, així com els militars i diplomàtics que treballin com a tals a l'estranger, els científics que treballin a bases espanyoles a l'estranger, els tripulants de vaixells i naus explotades per unitats residents a Espanya.

1.2. Àmbit geogràfic.

L'enquesta es realitza a tot el territori espanyol. A Ceuta i a Melilla s'efectua des del 2n trimestre de 1988. L'enquesta ha de donar estimacions desagregades fins el nivell de província.

1.3. Àmbit temporal.

L'EPA és una enquesta de periodicitat trimestral. Les entrevistes són realitzades al llarg de les tretze setmanes del trimestre.
La informació recollida fa referència a la setmana abans de la que es produeix l'entrevista, amb les excepcions que se citen tot seguit:

- Quatre setmanes anteriors en el cas dels mètodes de recerca de feina, de les peculiaritats de la jornada de feina i les relatives als estudis que s'han seguit.
- El diumenge de la setmana de referència per a l'edat. El mateix per a la inscripció en l'oficina de treball de l'administració.
- Dues setmanes posteriors al diumenge de la setmana de referència quan es demana per la disponibilitat per treballar
- Hi ha preguntes sobre el lloc de residència d'un any abans i, en el cas de les persones de 16 o més anys, sobre la seva situació respecte a l'activitat.
- L'any anterior, en la pregunta "Va realitzar alguna tasca remunerada, per compte propi o com aliè, en cap moment de l'any passat"?

En resum, el període de referència dels resultats de l'EPA és el trimestre i el període de referència de la informació és la setmana immediatament anterior (de dilluns a diumenge) a la de l'entrevista segons el calendari.

2. Marc

Per definir el **marc zonal** de l'EPA cal detallar la divisió administrativa d'Espanya. L'Estat Espanyol es divideix en 17 comunitats autònomes i dues ciutats autònomes. Aquestes es divideixen en províncies, fins un total de 50. Les ciutats autònomes no estan integrades en l'organització provincial, però es tractaran de manera similar, arribant a tenir-ne, doncs, una llista de 52. Les províncies es divideixen en municipis (consells, en el Principat d'Astúries i Galícia) i aquests, en districtes municipals, que es subdivideixen en **seccions censals**. Les seccions censals són utilitzades en tots els treballs de l'INE on és necessària una divisió inframunicipal, com ho és amb finalitats electorals: cada secció té una mida mínima de 500 electors i una mida màxima de 2000. Els seus límits estan perfectament definits. Hi ha unes 36000 seccions censals a Espanya, de les quals 5046 són a Catalunya (a 1 de gener de 2013). Es fa una actualització de les seccions censals el primer dia de cada any.
De cada secció censal triada per a l'enquesta, es té la relació d'habitatges familiars amb l'adreça postal. Aquest **marc d'habitatges** s'obté durant els recorreguts que es fan durant els treballs censals i s'actualitza periòdicament.

3. Tipus de mostreig. Les unitats mostrals

A cada província se selecciona una mostra independent. S'utilitza el tipus de mostreig comú a les enquestes de les llars: el mostreig bietàpic amb estratificació de les unitats de primera etapa. Aquestes unitats primàries són les seccions censals. Les unitats secundàries o de segona etapa són els habitatges familiars principals i els allotjaments fixos.
Aquest tipus de mostreig facilita les tasques de manteniment d'un marc actualitzat de selecció d'habitatges i, a més, redueix el temps de desplaçament dels entrevistadors i el temps necessari per la localització i l'accés a les llars.

3.1. Estratificació.

Per cada província, les unitats de primera etapa (seccions censals) s'agrupen en nou estrats, segons la importància demogràfica del municipi al qual pertanyen (criteri geogràfic).
Considerarem dos tipus de municipis: els autorepresentats i els corepresentats:

- Municipis autorepresentats: No poden estar representats per altres municipis de la mostra i, per tant, han de tenir representació mostral pròpia. Corresponen als estrats 1, 2 i 3, essent el primer la capital de província. L'estrat 2 correspon a municipis importants en relació a la capital i el 3, a municipis també de certa importància o amb una població superior als 100.000 habitants.
- Municipis corepresentats: Corresponen als estrats 4-9, segons la seva població:
  - Estrat 4: 50.000 - 100.000 habitants
  - Estrat 5: 20.000 - 50.000 habitants

- Estrat 6: 10.000 - 20.000 habitants
- Estrat 7: 5.000 - 10.000 habitants
- Estrat 8: 2.000 - 5.000 habitants
- Estrat 9: menys de 2.000 habitants

A la pràctica no totes les províncies tenen seccions en tots els estrats. En cas que hi hagi poca població en un, es pot fusionar amb un altre.

Aquesta definició dels estrats s'actualitza cada 10 anys amb la informació procedent dels Censos de Població.

3.2. Subestratificació.

Dins de cada estrat, les seccions s'agrupen en subestrats en funció de la categoria socioeconòmica de la seva població activa (criteri socioeconòmic). Es consideren dos grups de seccions:

- Seccions de municipis petits (estrats 7, 8 i 9): Es considera que la variabilitat és relativament petita respecte de les variables objectiu i, en qualsevol cas, ben explicada pel territori al qual pertanyen. Per tant, se'ls assigna la comarca on es localitzen per així distribuir la mostra en grups més homogenis i poder obtenir en un futur estimacions més desagregades.
- Seccions de municipis grans i mitjans (estrats 1-6): S'agrupen utilitzant tècniques d'anàlisi de conglomerats sobre variables socioeconòmiques procedents del Cens de població i habitatges 2001 i de l'Agència Estatal d'Administració Tributària (AEAT), que es consideren correlacionades amb les que són objectiu de l'enquesta. Les variables utilitzades són:
  - Relació amb l'activitat
  - Nivell d'estudis
  - Grups d'edat
  - Població estrangera
  - Condició socioeconòmica (deriva d'altres variables, procedeix del Cens)
  - Variable de renda

  En tractar-se de municipis més grans, ja tenien pràcticament garantida la representació mostral de la comarca, per la qual cosa s'ha utilitzat la informació auxiliar esmentada per a la creació de grups homogenis de secció i la conseqüent millora de la precisió de les estimacions.

Com a pas previ a l'anàlisi de conglomerats, s'han estandarditzat les variables dins de cada estrat amb mitjana 0 i desviació típica 1, llevat de les variables "percentatge d'aturats", "percentatge de joves", "renda per llar amb perceptors", "renda capital mobiliari i immobiliari sobre renda total" així com de "renda agrària sobre renda total", a les quals se'ls ha estandarditzat amb una desviació típica 2, per dotar-les d'una ponderació superior a la resta.

4. Grandària, afixació i selecció de la mostra

4.1. Grandària de la mostra.

Segons els càlculs de l'INE, es va establir una grandària mostral mínima de 3000 seccions i 20 llars per secció a tot l'Estat amb l'objectiu de donar estimacions

trimestrals a nivell de província sense superar els límits del pressupost. Degut a exigències de la Unió Europea i el desig de millora de la precisió d'algunes províncies, hi ha hagut canvis en les xifres, de tal forma que actualment la grandària mostral és de 3822 seccions i 18 habitatges per secció, llevat de les províncies que contenen les 5 ciutats amb més habitants de l'Estat, on es prenen unes 22 llars per secció.

4.2. Afixació de la mostra.

L'afixació provincial és de compromís entre la uniforme i proporcional: una part es reparteix per igual entre totes les províncies (ciutats autònomes incloses), mentre que l'altra part es reparteix de forma proporcional a la mida poblacional. D'aquesta manera, es pot donar estimacions a nivell de comunitat autònoma i fins i tot algunes de província perquè cap d'elles queda sense representació. S'imposa que les províncies i ciutats autònomes tinguin un nombre de seccions múltiple de 13 amb l'objectiu de distribuir-les al llarg del trimestre, que té 13 setmanes. En quant a l'afixació entre estrats, dins de cada província la mostra es reparteix entre els estrats i els subestrats de forma proporcional a la mida poblacional.

A Catalunya el nombre de seccions que se seleccionen és de 351, amb uns 20 habitatges per secció, gràcies a la mostra més gran presa a la província de Barcelona. Les seccions són repartides per província i estrat com es mostra a la taula 3.1.

Taula 3.1. Seccions censals per província i estrat

	E1	E2	E3	E4	E5	E6	E7	E8	E9	Total
Barcelona	55	0	33	19	21	12	10	3	3	156
Girona	15	0	0	0	19	12	10	13	9	78
Lleida	15	0	0	0	0	5	3	6	10	39
Tarragona	19	12	0	0	12	9	9	9	8	78
Total	104	12	33	19	52	38	32	31	30	351

4.3. Selecció de la mostra.

Les unitats de primera etapa se seleccionen amb probabilitat proporcional al nombre d'habitatges familiars. Dins de cada secció s'escullen unes 18 unitats de segona etapa (una mica més a Catalunya) amb probabilitat igual. En ambdós casos, s'usa un mostreig sistemàtic amb ordenació prèvia segons subestrats, en el cas de les seccions censals, i segons illes, en el cas de les llars. S'entrevisten tots els individus de cada llar, però dels menors de 16 anys només es recullen dades de caràcter demogràfic.

La probabilitat de selecció d'una llar i que pertany a la secció j d'un estrat h, on s'han afixat K_h seccions, seria:

$$P(V_{ijh}) = P(S_{jh}) \cdot P(V_{ijh}|S_{jh}) = K_h \cdot \frac{V_{jh}}{V_h} \cdot \frac{m}{V_{jh}} = \frac{K_h \cdot m}{V_h} = \frac{v_h^t}{V_h},$$

on:

- $P(S_{jh})$: Probabilitat de selecció de la secció j de l'estrat h.
- $P(V_{ijh}|S_{jh})$: Probabilitat de selecció de l'habitatge i condicionada a la selecció de la secció j.
- V_{jh}: Total de domicilis de la secció i.
- V_h: Total de domicilis de l'estrat h.

- m: Nombre fix de llars seleccionades per secció. En la metodologia documentada de l'EPA, s'informa que $m = 18$.
- v_h^t: llars de la mostra teòrica a l'estrat h.

5. La mostra al llarg del temps

5.1. Renovació parcial de la mostra.

La mostra es renova d'aquesta manera:

- Unitats primàries: La mostra de seccions censals roman en el temps per rendibilitzar el treball de camp realitzat en quant al directori d'habitatges, plànols i guies de carrers, i en el coneixement del veïnat per part dels entrevistadors i viceversa. Tan sols es modifica en el cas que s'esgotin els domicilis enquestables de la secció, quan s'actualitzen les probabilitats mostrals o quan, en funció d'informació externa, es modifica el repartiment de la mostra per estrats, és a dir, l'afixació.
- Unitats secundàries: Cada trimestre es renova 1/6 part de les llars de les seccions censals, mitjançant els **torns de rotació**. D'aquesta forma, quan una llar ha estat entrevistada durant sis trimestres consecutius, deixa de col·laborar en l'enquesta i és substituïda per una altra. En vista a les renovacions del trimestre següent, l'entrevistador recorre la secció i incorpora al marc les noves construccions i els habitatges que han deixat d'estar buits. És per tant una enquesta contínua de panel rotant 1/6.

5.2. Distribució de la mostra en el temps.

La distribució de la mostra és uniforme en el temps, el que implica que a cada província el nombre de seccions per estrat i torn de rotació de cada setmana és semblant. Així es pretén obtenir una representació adequada de la població en quant a les variables objectiu al llarg del període de referència de l'enquesta, que, com ja s'ha dit, és d'un trimestre. Cadascuna de les seccions de la mostra és visitada en una de les 13 setmanes d'aquest període.

La importància d'obtenir una mostra temporalment ben distribuïda, de forma que es tingui una representació adequada de tot el període temporal, rau en que una unitat mostral pot representar valors diferents de les variables a estimar al llarg del temps. Per posar un exemple, el nombre d'aturats que hi ha a una llar pot canviar d'una setmana per l'altra.

5.3. Actualitzacions en el marc de l'enquesta.

Les variacions que es donen a la població de forma continuada, tant en termes de característiques com de la distribució espacial, exigeixen de realitzar actualitzacions en els marcs, que repercuteixen en l'estructura mostral.

A l'EPA es consideren les actualitzacions següents:

- En el marc d'habitatges: Tenen caràcter restringit a les seccions de la mostra. Es tracta en incorporar les noves llars.
- En el marc de seccions censals: Són conseqüència de variacions en les unitats primàries seleccionades per a la mostra. Aquestes modificacions poden ser degudes a:
  - Partició de seccions

  - Fusió de seccions
  - Reajustament de límits de seccions
- De caràcter general: Relativa a totes les seccions censals i llars, quan es realitzen els Censos de població i habitatges i es disposta per tant, de nova informació. Es revisa la definició dels estrats i dels subestrats i s'actualitza la probabilitat de selecció de la selecció.
- De les probabilitats d'inclusió del seleccionat: Es pretén que, amb el mínim possible de canvis, la mostra de seccions sigui equivalent a una mostra seleccionada l'any de l'actualització. Es fa cada tres anys.

Capítol 4
Obtenció dels estimadors: metodologia actual

1. Introducció

Quan treballem amb enquestes dirigides a les llars, una de les dificultats amb les quals es fàcil topar-nos és amb no poder contactar amb cap membre d'un domicili, malgrat que els entrevistadors repeteixen les visites i les trucades. Pot passar amb més freqüència a les llars on hi viu només una persona o on hi viuen parelles on tots dos treballen. Això pot causar la sobrerepresentació de la gent d'edat avançada. Per corregir-ho i així garantir una bona representitivitat de la mostra, l'INE aplica tècniques de reponderació on es rectifiquen els pesos o factors d'elevació originals deduïts del disseny de l'enquesta de manera que es reflectirà correctament el pes que té cada grup (segons les variables que es tinguin en compte) a la població original, obtinguda de fonts externes fiables, com ho és el Cens de població i habitatges o el Padró municipal d'habitants.
Per tant, per dur a terme un procés de reponderació és necessari triar unes variables explicatives o auxiliars que siguin presents tant a l'enquesta com a la font estadística externa a aquesta.
S'han utilitzat les variables auxiliars següents:

- La població de 16 anys i més per grups d'edat i sexe, a nivell de comunitat autònoma.
- Població de 16 anys i més per comunitat autònoma i nacionalitat -espanyola o estrangera-.
- Població de 16 anys i més per província.
- Població menor de 16 anys per grups d'edat i sexe, a nivell de comunitat autònoma.
- Població menor de 16 anys per província.

A partir del 2014, l'INE aprofita el canvi que suposa l'actualització de les poblacions de referència provinents del Cens de població i habitatges de 2011, per incorporar com a variables auxiliars per al procés, la mida de la llar, segons les persones que hi resideixen, i la població a cada província per sexe i edat (16-29 anys, 30-49 anys, 50 anys i més).

2. Metodologia per l'obtenció dels estimadors

De forma intuïtiva es pot establir que el factor d'elevació depèn de la probabilitat que una unitat pertanyi a la mostra, de que aquesta unitat faciliti o no la informació sol·licitada, i dels ajustos a fonts externes que minvin la variabilitat de les estimacions.

Atenent a aquesta idea, per a l'obtenció de les estimacions se seguirà el procediment habitual utilitzat en les enquestes demogràfiques [**14**]:

- Pas 1. Estimador no esbiaixat d'expansió (Horvitz-Thompson, H-T): Assignant-li a cada unitat mostral un pes de disseny que sigui la inversa de la probabilitat de pertànyer a la mostra, es compensa les desiguals probabilitats de selecció i s'obté un estimador sense biaix.
- Pas 2. Correcció de falta de resposta: Malauradament, la mostra efectiva (de les unitats que han col·laborat en l'estudi) i la teòrica no coincideixen. Es pot corregir el biaix de la manca de resposta total repartint el pes de les unitats que no han respost entre les que sí que ho han fet, de manera que la suma d'aquests pesos de la mostra efectiva coincideixi amb la dels pesos del disseny de la mostra teòrica, obtinguts al pas anterior.
- Pas 3. Reponderació mitjançant tècniques de calibratge amb fonts externes.

El resultat d'aquest procés és un pes o factor d'elevació per a cada element de la mostra efectiva.

2.1. L'estimador de Horvitz-Thompson.

L'estimador H-T utilitza com pesos la inversa de les probabilitats d'inclusió a la mostra (Capítol 2, Secció 4.3). Si anomenem Y a la variable objectiu, l'expressió serà:

$$\hat{Y}_{H-T} = \sum_h \frac{V_h}{v_h^t} \cdot \sum_{i \in h} y_i \tag{1}$$

2.2. Correció per manca de resposta.

Es farà estimant la probabilitat que les unitats responguin i incloent-la dins la fórmula de l'estimador. Si considerem que dins de cada estrat la probabilitat de resposta P_{Rh} és igual per totes les unitats, es té que:

$$P_{Rh} = \frac{V_h}{v_h^t}, \tag{2}$$

on v_h és la mostra efectiva de llars a l'estrat h. Incorporant la inversa d'aquesta probabilitat a l'estimador, s'obté:

$$\hat{Y}_{H-TCorr} = \sum_h \frac{V_h}{v_h^t} \cdot \frac{v_h^t}{v_h} y_i = \sum_h \frac{V_h}{v_h} \sum_{i \in h} y_i = \sum_h \hat{Y}_{H-TCorr(h)} \tag{3}$$

2.3. Tècniques de calibratge.

Primer es farà servir un estimador de raó separat que pren com a variables auxiliars la població de 16 i més anys referida a la meitat del trimestre, P_h. Aquest estimador

té l'expressió:

$$\hat{Y}_{Cal1} = \sum_h \frac{\hat{Y}_{H-TCorr(h)}}{\hat{P}_{H-TCorr(h)}} \cdot P_h \tag{4}$$

Llavors, si substituïm el numerador i el denominador de (4) per l'expressió resultant a (3), obtenim:

$$\hat{Y}_{Cal1} = \sum_h \frac{\frac{V_h}{v_h} \sum\limits_{i \in h} y_i}{\frac{V_h}{v_h} \sum\limits_{i \in h} p_i} \cdot P_h = \sum_h \frac{P_h}{p_h} \cdot y_h = \sum_S d_k \cdot y_k \tag{5}$$

Posteriorment, s'ajustarà per les fonts externes. S'utilitzen les variables auxiliars enunciades a la introducció del capítol que fan referència a la població de 16 anys o més.
Anomenem xj a cadascuna de les p variables auxiliars i X_j al total conegut a la comunitat autònoma, és a dir:

$$X_j = \sum_{k \in U} x_{jk} \tag{6}$$

on U és la població. Segurament amb els factors que tenim fins ara, $d_k = P_h/p_h$, la mostra no serà equilibrada:

$$\hat{X}_j \neq \sum_{k \in s} d_k x_{jk} \tag{7}$$

L'objectiu del calibratge serà obtenir uns nous pesos wk que, allunyant-se el menys possible dels anteriors dk, verifiqui les equacions d'equilibri:

$$\hat{X}_j = \sum_{k \in s} w_k x_{jk} \tag{8}$$

En altres paraules, pretenem obtenir uns nous factors d'elevació que ens permetin estimar els totals poblacionals coneguts correctament i que s'allunyin el mínim possible dels de H-T, que tenen la propietat de ser no esbiaixats.
Per tant, això es transforma en un problema d'optimització. La seva resolució requereix de la definició d'una distància per establir una diferència entre els pesos originals i els nous, i de resoldre un problema de programació.

L'INE parteix d'un estimador amb bones propietats, com podria ser el de H-T. Utilitza per resoldre'l el mòdul CALMAR de SAS, que resol en aquest cas un problema de programació lineal, que es planteja així:

$$\min_{w_k} \sum_{k \in s} d_k G(\frac{w_k}{d_k})(k = 1, ..., n)$$

s.a.:

$$\begin{cases} \sum_{k \in s} w_k x_k = X \\ x'_k = (x_{k1}, ..., x_{kJ}) \\ X' = (X_1, ..., X_J), \end{cases}$$

on $G(X)$ és una funció positiva i estrictament convexa i que compleix G(1)=G'(1)=0 i G"(1)=1, per tal de garantir l'existència de solució.

L'INE empra el mètode lineal, la funció de distància de la qual és:

$$G(X) = \frac{1}{2}(x-1)^2.$$

Algunes institucions europees utilitzen la funció logit, en lloc de la lineal. El CALMAR també dóna l'opció d'utilitzar, a més, altres funcions de distància, com el *raking ratio*.

3. Els errors mostrals

S'originen dos grans tipus d'errors, uns en obtenir la mostra i uns altres que en són aliens a aquesta i que poden afectar fins i tot a un cens (per exemple, un entrevistador que s'equivoca, un enquestat que no dóna una informació correcta o uns errors de depuració de les dades). L'objectiu del treball és evidentment tractar-ne els primers. La complexitat dels dissenys d'enquestes com l'EPA, un clar exemple d'enquesta complexa, però, fa inviable obtenir una fórmula directa de la variància (i per extensió, de l'error mostral).

3.1. Metodologia de l'INE.

El mètode triat per l'INE per estimar els errors de mostreig de l'EPA és el de semimostres reiterades, basant-se en l'ús que la U.S. Census Bureau (EUA) en fa d'aquesta tècnica.
Es duu a terme partint tota la mostra de PSUs en dues submostres disjuntes. Això es repeteix vint cops i, conseqüentment, s'obtenen 40 semimostres.
La variància s'estima utilitzant l'expressió:

$$\hat{V}(\hat{X}) = \frac{1}{40}\sum_{i=1}^{40}(\hat{X}_i - \hat{X})^2, \tag{9}$$

on X és la variable objectiu i el subíndex i fa referència a l'estimació obtinguda amb la semimostra i.

3.2. Metodologia d'altres instituts d'estadística.

Arreu del món, sembla que el Jackknife, la linealització de Taylor i el bootstrap són els mètodes més utilitzats:

- Alemanya: Linealització de Taylor.
- Austràlia: Jackknife per grups. Fins el 2002, semimostres
- Canadà: Jackknife
- Colòmbia: Linealització de Taylor
- França: Fórmules teòriques
- Itàlia: Estimador de regressió generalitzat
- Mèxic: Linealització de Taylor
- País Basc: Linealització de Taylor
- Països Baixos: Bootstrap
- Portugal: Jackknife
- Polònia: Bootstrap
- Regne Unit: Linealització de Jackknife
- USA: "Metodologia similar a la Bootstrap"(sic)

Part 3

Proposta metodològica

Capítol 5
Descripció i tractament de les variables

1. Selecció de variables

Partim de bases de dades trimestrals facilitades per l'INE, que ja estan degudament tractades. Actualment, el fitxer de treball compta amb 175 variables, mancant-hi els torns de rotació on pertanyen les llars aquell trimestre i la informació sobre la falta de resposta.
Per a la realització d'aquest treball, hem agafat els quatre trimestres del 2013 i els del 2014. Així tindrem representats dos exercicis sencers amb les seves fluctuacions trimestrals habituals i dos mètodes de calibratge diferents, com comentàvem al Capítol 4.1. Per a la manipulació de les dades, s'han utilitzat les onze variables següents:

- **PROV**: Fa referència a la província:
  - 08: Barcelona
  - 17: Girona
  - 25: Lleida
  - 43: Tarragona
- **STRAT**: És l'estrat (1-9) al que correspon l'element de la mostra.
- **CENS**: És l'identificador de la secció censal de la mostra, per estrat i província. El seu rang de valors va del 001 fins al nombre màxim de seccions de la mostra a cada estrat (i per a cada província).
- **HId**: Es tracta de l'identificador de la llar. Comença per 1 i el valor més alt és el nombre d'habitatges que hi ha a la mostra final a Catalunya.
- **PId**: És el nombre assignat a cada individu de la llar.
- **AGE**: Pren el valor de l'edat que té l'individu el diumenge de la setmana de referència.
- **SEX**: És el sexe de la persona (1: Home, 6: Dona)
- **SIT**: Classifica els individus de 16 anys o més en inactius i en actius (ocupats o aturats):
  - 03: Ocupats subempleats per insuficiència d'hores
  - 04: Resta d'ocupats
  - 05: Aturats que busquen la primera feina

  - 06: Aturats que ja havien treballat abans
  - 07: Inactius 1 (desanimats -actius potencials-)
  - 08: Inactius 2 (altres actius potencials)
  - 09: Inactius 3 (resta d'inactius)
- **NAT**: És la nacionalitat (1: Espanyola o doble nacionalitat, 6: No espanyola)
- **FACT**: És el factor d'elevació calculat per l'INE.

Taula 5.1. Mostra hipotètica de variables triades

	PROV	STRAT	CENS	HId	PId	AGE	SEX	SIT	NAT	FACT
1	8	1	1	1	1	46	1	4	1	549.29
2	8	1	13	260	2	35	1	6	1	495.44
3	8	3	6	1331	3	19	1	5	1	525.23
4	8	4	10	2151	2	64	1	4	1	1021.24
5	17	1	7	3564	1	38	1	4	1	166.18
6	17	5	18	4003	2	61	6	9	1	181.2
7	17	9	3	4653	1	53	1	4	1	261.51
8	25	8	5	5187	1	88	6	9	1	229.5
9	43	5	5	5961	1	40	1	4	1	260.6
10	43	9	3	6589	1	72	1	9	1	245.23

2. Transformació de variables

Hem creat noves variables a partir de les citades abans, per tal de poder realitzar les tasques de manipulació que explicarem més endavant. Es tracta de les següents:

- **SIZE**: És el nombre de persones que hi ha a la llar. La categoria 5 correspon a 5 o més.
- **gAGE**: És una categorització de la variable AGE, en intervals de 5 anys, a excepció del primer, que correspon al grup d'edat de 16 a 19 anys i del darrer, que correspon al grup de 65 anys o més:
  - 0: 16-19 anys
  - 1: 20-24 anys
  - 2: 25-29 anys
  - 3: 30-34 anys
  - 4: 35-39 anys
  - 5: 40-44 anys
  - 6: 45-49 anys
  - 7: 50-54 anys
  - 8: 55-59 anys
  - 9: 60-64 anys
  - 10: 65 anys i més
- **gAGE2**: És una categorització de la variable AGE, amb només tres nivells:
  - 0: 16-29 anys
  - 1: 30-49 anys
  - 2: 50 anys i més

- **SEXgAGE**: És una combinació de les variables SEX i gAGE, amb un total de 22 categories. Un individu amb valor 6.9 en aquesta variable representa una dona d'una edat compresa entre els 60 i 64 anys.
- **SEXgAGE2**: És també una unió entre les variables SEX i gAGE2. Una persona amb valor 1.0 correspondria a un home d'entre 16 i 29 anys.
- **AGE4**: És una agrupació de la variable AGE feta per publicar els resultats pels grups d'edat següents:
  - 0: 16-19 anys
  - 1: 20-24 anys
  - 2: 25-54 anys
  - 3: 55 anys i més
- **EDUC**: És el nivell de formació assolit, seguint la classificació CCED-2000 fins el 2013 i després, la CNED-2014:
  - 1: Analfabets i educació primària
  - 2: Educació secundària 1a. etapa
  - 3: Educació secundària 2a. etapa
  - 4: Educació superior
- **ACTIN**: Es tracta d'una variable binària que pren valor 1 si l'entrevistat de 16 anys o més forma part de la població activa i 0, altrament.
- **EMPUN**: Pren valor 1 si l'individu està ocupat, valor 2 si està aturat i 0 si és inactiu.
- **UNEMP**: Pren valor 1 si l'individu està desocupat, valor 0 si està ocupat i NA si és inactiu.
- **EMPOP**: És com la variable EMPUN, però la categoria 0 inclou també els aturats (abans la 2).
- **CENSOR**: És la secció censal. Han estat ordenades segons aquest criteri (en aquest ordre):

PROV, STRAT, CENS

- **ECO**: És el sector econòmic:
  - Agric: Agricultura
  - Indus: Indústria
  - Cons: Construcció
  - Serv: Serveis
- **BRAN**: És la branca d'activitat:
  - 1: Agricultura, ramaderia, silvicultura i pesca
  - 2: Indústries extractives, energia, aigua i residus
  - 3: Alimentació, tèxtil, fusta, paper i arts gràfiques
  - 4: Química i cautxú
  - 5: Metal·lúrgia
  - 6: Maquinària, material elèctric i de transport
  - 7: Construcció
  - 8: Comerç engròs i reparació vehicles motor i motocicletes
  - 9: Comerç detall
  - 10: Transport i emmagatzematge
  - 11: Hostaleria
  - 12: Informació i comunicacions
  - 13: Activitats financeres i assegurances
  - 14: Activitats immobiliàries, professionals i tècniques

- 15: Activitats administratives i serveis auxiliars
- 16: Administració pública
- 17: Educació
- 18: Sanitat i serveis socials
- 19: Activitats culturals i esportives, i altres serveis

TAULA 5.2. Mostra hipotètica de noves variables I

	SIZE	gAGE	SEXgAGE	gAGE2	SEXgAGE2	AGE4	EDUC
1	3	6	1.6	1	1.1	3	3
2	3	4	1.4	1	1.1	3	1
3	4	0	1	0	1	1	2
4	2	9	1.9	2	1.2	4	2
5	1	4	1.4	1	1.1	3	2
6	4	9	6.9	2	6.2	4	2
7	3	7	1.7	2	1.2	3	2
8	1	10	6.1	2	6.2	4	2
9	4	5	1.5	1	1.1	3	2
10	2	10	1.1	2	1.2	4	1

TAULA 5.3. Mostra hipotètica de noves variables II

	ACTIN	EMPUN	UNEMP	EMPOP	CENSOR	ECO	BRAN
1	1	1	1	1	1	Serv	19
2	1	2	2	0	13	NA	NA
3	1	2	2	0	61	NA	NA
4	1	1	1	1	98	Serv	9
5	1	1	1	1	163	Serv	10
6	0	0	NA	0	189	NA	NA
7	1	1	1	1	228	Indus	3
8	0	0	NA	0	262	NA	NA
9	1	1	1	1	309	Cons	7
10	0	0	NA	0	346	NA	NA

3. Descripció de les dades

Presentem una bateria de representacions gràfiques amb comentaris, si escau, per tal d'explicar breument característiques de la mostra (i de la població).

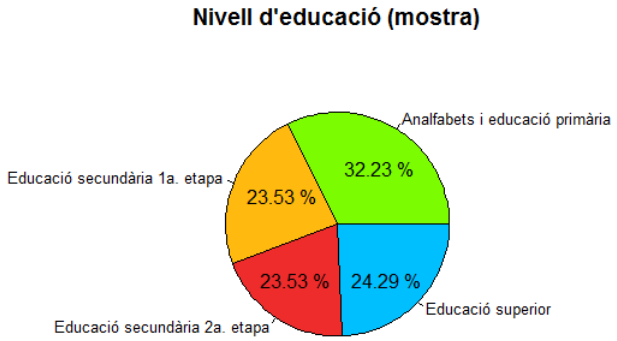

FIGURA 5.1. Màxim nivell educatiu assolit dels individus de la mostra
En temes de nivell acadèmic assolit, la població catalana es mostra bastant ben repartida entre els 4 grups.

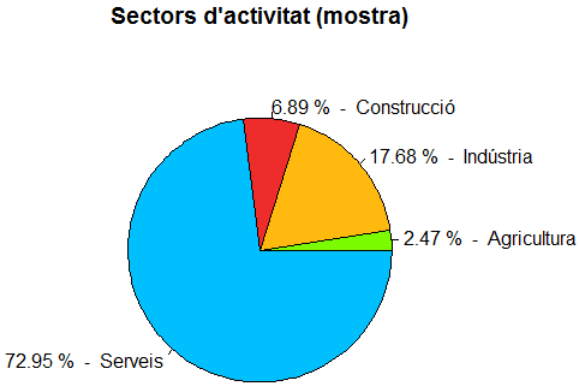

FIGURA 5.2. Sector d'activitat dels individus de la mostra
Pel que fa als sectors d'activitat, és conegut que la població catalana es dedica de forma molt majoritària al sector dels serveis, i que el sector primari representa una minoria, com passa als països desenvolupats.

Observem (FIGURES 5.3 i 5.4) que la població d'edat més avançada està sobrerepresentada a la mostra efectiva (tal com havíem introduït al Capítol 4, tot justificant la reponderació), i en major mesura en les dones grans. Els enquestadors de l'EPA visiten la llar i intenten que siguin tots els membres de 16 o més anys els que responguin, però en cas que no sigui possible, un altre membre els donarà la informació (és el que es coneix com una entrevista *PROXY*: és bastant fiable, ja que hi ha

poques preguntes qualitatives). És més senzill localitzar la gent gran a casa i dins d'aquest col·lectiu, les dones (vídues, mestresses de casa...).

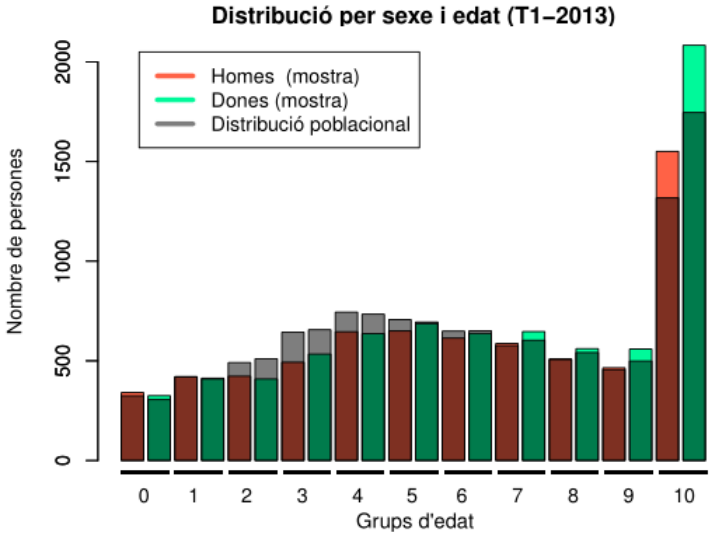

FIGURA 5.3. Distribucions mostral i poblacional, per sexe i edat (2013)
Veiem el nombre de persones a la mostra efectiva, per sexe i edat. Enfosquit, i barrejant-se amb els altres colors, veiem la distribució que hauria de tenir per ser igual que la poblacional. Aquesta s'ha obtingut a partir de la població per sexe i edat, aconseguida amb els factors, i dividint-la entre el factor d'elevació global.

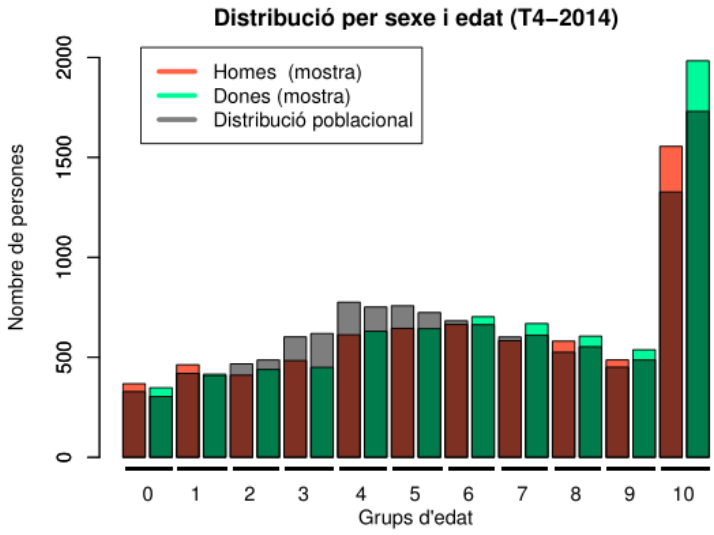

FIGURA 5.4. Distribucions mostral i poblacional, per sexe i edat (2014)
Passa el mateix que el 2013 (cosa que justifica que mostrem normalment les dades i els resultats d'un sol trimestre al llarg d'aquest treball).

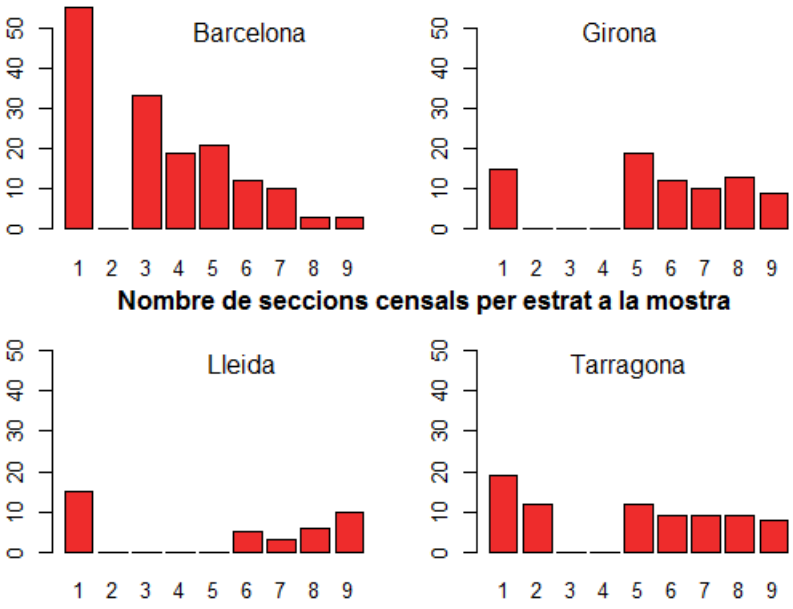

FIGURA 5.5. PSUs per estrat i província

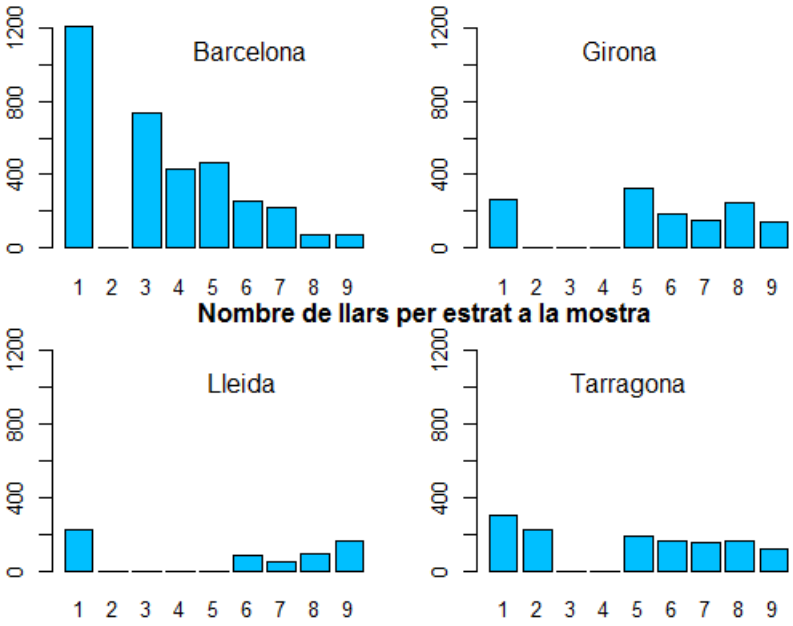

FIGURA 5.6. SSUs per estrat i província

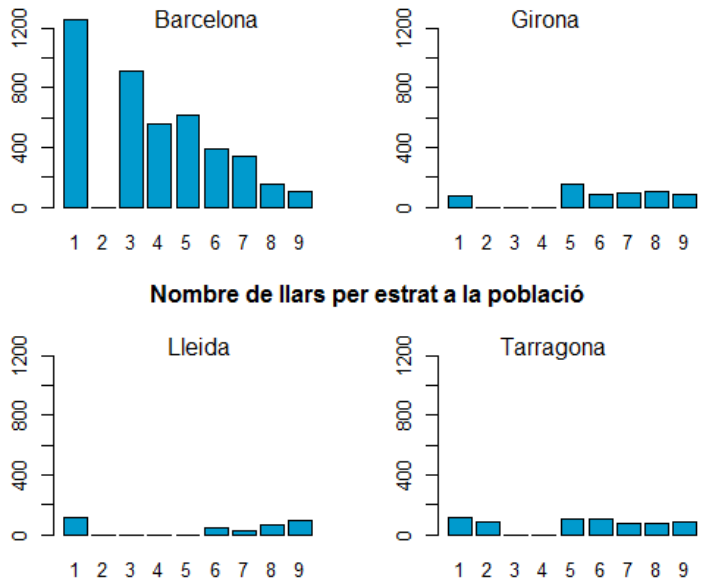

FIGURA 5.7. Llars per estrat i província a la població (en milers)
S'ha obtingut amb els factors de l'INE. L'estrat 1 de Barcelona sembla estar sobrerepresentat (quan ja té una mostra prou gran, en relació amb els altres) i, per exemple, passa el contrari amb l'estrat 8 de la mateixa província.

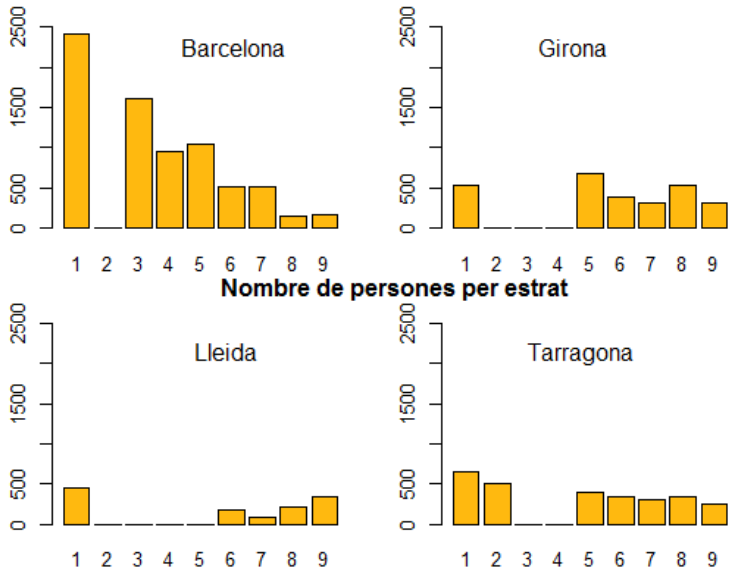

FIGURA 5.8. Persones a la mostra efectiva, per estrat i província

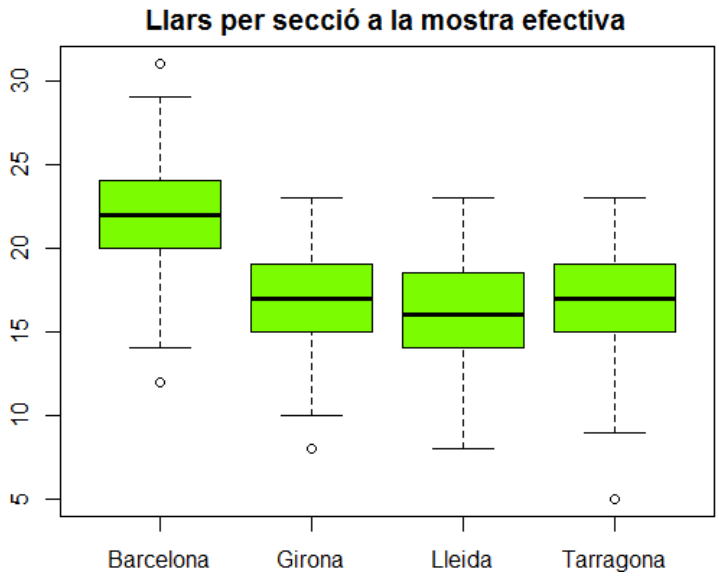

FIGURA 5.9. Llars per secció, a cada província
Com sabíem, es prenen més llars per secció a la província de Barcelona.

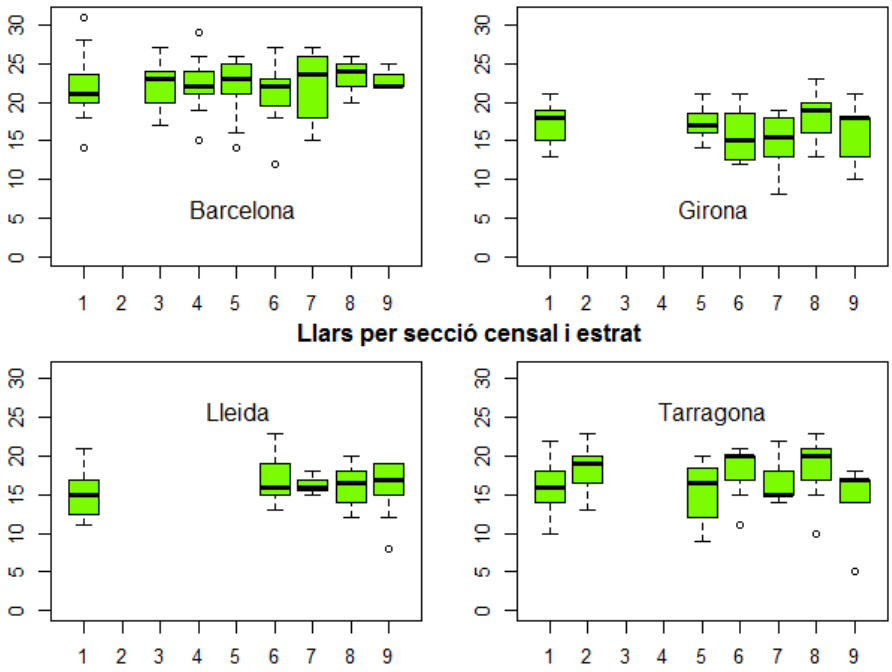

FIGURA 5.10. Llars per secció, a cada província i estrat
Crida l'atenció que en certs casos, la mostra presa d'una secció ha estat de menys de 10 llars.

Capítol 6
Estimació de les principals magnituds de treball i del seu error

En aquest capítol, mostrarem els diferents estimadors que hem obtingut per a les nostres variables d'interés, centrant-nos en les estimacions dels totals. Primerament, explicarem com hem fet els càlculs mitjançant R, per posteriorment presentar la metodologia que hem seguit per calcular cada factor. També hem seguit la plantejada per Deville, Särndal i Sautory (1993)[**14**], explicada al Capítol 4. Considerem, així l'estimador de Horvitz-Thompson, el de raó separat i el calibrat. N'havíem considerat d'altres en un principi, com també explicarem. Addicionalment, disposem dels resultats publicats per l'INE, així com els factors que han utilitzat, l'obtenció dels quals ha estat detallada al Capítol 4. Hem fet totes les estimacions per al primer trimestre de 2013.

1. El paquet *survey* d'R

Un dels grans avantatges que té l'R és que es tracta d'un software totalment lliure. A més a més, gràcies a l'ús i col·laboració de l'extensa comunitat d'usuaris, la qualitat del producte és comparable a la d'altres de pagament.
En el cas del procés de calibratge, i com s'ha comentat abans, l'INE utilitza CALMAR, un mòdul de SAS. Però degut a que SAS es tracta d'una eina costosa en termes econòmics, l'R ens ha semblat una alternativa interessant, i més en èpoques d'austeritat. El paquet survey, creat per Thomas Lumley, conté nombroses funcions per a l'anàlisi d'enquestes complexes. Malgrat que una de les tasques que hem realitzat en aquest treball és la programació dels mètodes de remostreig, finalment hem optat per l'ús de les funcions implementades en aquesta llibreria d'R perquè integra totes les eines necessàries per al tractament d'enquestes complexes i amb un cost computacional molt raonable. N'expliquem breument aquelles que ens han servit per al desenvolupament del treball.

1.1. Funció svydesign.
Serveix per especificar el disseny complex de l'enquesta i per treballar amb les funcions d'aquest paquet.

L'hem utilitzat amb aquests paràmetres: svydesign(ids,strata,data,weights,nest), on:

- **ids** són les unitats mostrals, començant per les més grans (primàries),
- **strata** és la variable o variables d'estratificació,
- **data** és la taula de dades,
- **weights** són els pesos mostrals,
- **nest**, si pren valor TRUE, reetiqueta els identificadors clúster per fer complir l'anidació dins dels estrats.

1.2. Funció as.svrepdesign.

Aquesta funció ens han estat d'ajut per estimar les variàncies mostrals mitjançant els diferents mètodes de remostreig. L'hem cridat mitjançant el codi següent: as.svrepdesign(desing,type,hadamard.matrix,replicates,mse,compress,small,large), on:

- **design** és el disseny svydesign,
- **type** és el mètode de replicació ("BRR", "bootstrap"...),
- **hadamard.matrix** és la matriu Hadamard per al mètode BRR,
- **replicates** és el nombre de replicacions,
- **mse**, si és igual a TRUE, calcularà l'EQM,
- **compress**, si pren valor FALSE, no utilitzarà una representació comprimida de la matriu de ponderacions replicades,
- **small** i **large**, serveixen per especificar com s'utilitzen en el BRR els estrats amb un o més PSUs, respectivament.

1.3. Funcions surveysummary: svytotal, svymean, svyvar i svyratio.

S'utilitzen per calcular totals, mitjanes, variàncies i ràtios, respectivament. L'hem fet funcionar així: svy***(x,design,deff), on:

- ******* fa referència a si volem mitjana (mean), total, etc.,
- **x** són les dades,
- **design** és el disseny svydesign o svrepdesign,
- **deff**, si pren valor TRUE, indica que desitgem que ens mostri l'efecte del disseny, el DEFF.

1.4. Funció svytable.

L'emprem per fer taules de contingència amb les dades del disseny. Hem seguit l'estructura: svytable(formula,design), on:

- **formula** és la fórmula del model, especificant les marginals de la taula,
- **design** és el disseny svydesign.

1.5. Funció calibrate.

L'hem implementat d'aquesta forma: calibrate(design,formula,population,calfun, bounds,aggregate), on:

- **design** és el disseny svydesign o svrepdesign,
- **formula** és la fórmula del model de calibratge,
- **population** és el vector amb els totals poblacionals,
- **calfun** és la funció de calibratge (utilitzarem la lineal, l·linear"),
- **bounds** són els límits dels pesos del calibratge (obligatoris per a funcions logit),

- **aggregate** (o aggregate.stage) especifica quina és l'etapa del mostreig a la qual els totals fan referència. En el nostre cas és la segona (SSU), ja que assignem un factor per llar.

1.6. Funció svyby.

Ens ha estat útil per construir els resums estadístics per nivells d'un factor. L'ús és senzill: svyby(formula/x, by ,design,...), on:

- **formula/x** és la fórmula on especifiquem les variables per passar-li a la funció, o bé són les dades,
- **by** és la fórmula o llista de factors que defineixen els subconjunts,
- **design** és el disseny svydesign o svrepdesign, que pot ser calibrat.

2. Estimadors no calibrats

2.1. L'estimador de Horvitz-Thompson.

És l'estimador dels pesos del disseny (factors d'elevació). És no-esbiaixat i es basa en les probabilitats de pertànyer a la mostra (Capítol 3.4.3). En el nostre cas, com més baixa sigui aquesta, a més habitatges haurà de representar, essent inversament proporcional (Capítol 4.2.1). Necessitem saber el nombre de llars poblacionals per cada estrat i el nombre d'habitatges per estrat a la mostra teòrica. Prendrem el valor 18, tot i que a Barcelona i a les grans províncies espanyoles, aquest nombre augmenta. Al principi de plantejar aquest treball, utilitzàvem diferents estimadors, tenint com a criteri el nombre de llars per estrat (podia ser 18, la mitjana real, o la mitjana per estrat de la mostra efectiva). Les vam descartar perquè hem de partir de la base que no tenim informació prèvia de la no-resposta.

L'INE fa estimacions trimestrals sobre la població a partir de diverses fonts, com ho són les dades obtingudes del Padró municipal. A partir dels factors d'elevació facilitats a les seves bases de dades, nosaltres podem conèixer aquests valors referits a la població. Tots els membres de la llar de 16 anys o més comparteixen el mateix factor: les unitats secundàries són les llars i no els individus que hi viuen. És per aquest motiu que només ens cal sumar els factors de tots els habitatges de la població que ens interessi, i d'aquesta manera obtindrem quantes persones hi viu a la població real, segons les dades de l'INE. Executant les instruccions en R (consultables a l'Annex D), obtenim les taula següent, per al primer trimestre de 2013:

Taula 6.1. Totals de l'activitat econòmica amb l'estimador H-T

T1-2013	**total**	**SE**	**2.50%**	**97.50%**	**CV(%)**	**deff**
Actius	3,849.1	65.45	3,720.9	3,977.4	1.70	5.4912
Ocupats	2,939.0	56.79	2,827.7	3,050.3	1.93	4.1152
Aturats	910.1	33.26	844.9	975.3	3.65	2.9683
At. no treb.	84.5	9.80	65.3	103.7	11.60	2.4306
At. sí treb.	825.6	29.89	767.0	884.2	3.62	2.6056
Inactius	2,881.65	53.41	2,776.96	2,986.33	1.85	3.6568

2.2. L'estimador de raó estratificat separat.

L'obtindrem a partir de considerar els habitants dels estrats a cada província i dividint-lo entre els que tenim a la mostra. És un estimador que presenta biaix. Per corregir-lo, podem recórrer al calibratge.

TAULA 6.2. Totals de l'activitat econòmica amb l'estimador de raó separat

T1-2013	total	SE	2.50%	97.50%	CV(%)	deff
Actius	3,409.8	56.04	3,300.0	3,519.7	1.64	5.1411
Ocupats	2,603.3	48.70	2,507.9	2,698.7	1.87	3.8614
Aturats	806.5	28.15	751.4	861.7	3.49	2.7116
At. no treb.	74.3	8.26	58.1	90.5	11.12	2.2198
At. sí treb.	732.2	25.33	682.6	781.9	3.46	2.3855
Inactius	2,548.2	45.60	2,458.8	2,637.6	1.79	3.4035

3. L'estimador calibrat. Mètodes de remostreig per estimar la variància

Els estimadors calibrats, creats per Deville i Särndal (1992)[**13**], són aquells que utilitzen pesos calibrats, els quals han de complir les condicions següents:

- Estar el màxim de prop dels pesos mostrals originals, d'acord amb una mesura de distància.
- Satisfer un conjunt de restriccions. Aquesta es refereix a la informació auxiliar (provinent de censos, registres administratius...), i en determinats casos, a l'interval en el que han d'estar els nous ponderadors.

Calibrarem a partir dels factors obtinguts amb l'estimador de raó separat i no pas amb el de H-T: considerem que la raó corregeix en certa mesura el biaix de la no-resposta i, per tant, serà un millor punt de partida per al calibratge[**33**].

Utilitzarem la funció lineal, comentada al Capítol 4.2.3.

Les estimacions de l'error les farem mitjançant la linealització de Taylor i tres mètodes de remostreig, introduïts al Capítol 2.4, descartant també els mètodes analítics, que són complexos i porten a adoptar assumpcions que no compleixen. En tot cas, pels estadístics que necessitem calcular, els mètodes que utilitzarem són vàlids[**3**].

Trobem necessari citar que no hem utilitzat la correcció per a la població finita perquè l'afectació que podria tenir en les estimacions, i tenint en compte que tenim una mostra gran, és menyspreable[**12**].

3.1. Estimació de la variància amb linealització de Taylor.

És el mètode que el paquet *survey* utilitza per defecte. Així doncs, només necessitarem executar les instruccions calibrate i svytotal, per tal d'obtenir la TAULA 6.3. Els totals estaran representats sempre en milers i els coeficients de variació, en percentatge.

TAULA 6.3. Totals de l'activitat econòmica amb raó l'estimador calibrat

T1-2013	total	SE	2.50%	97.50%	CV(%)	deff
Actius	3,675.7	19.9	3,636.6	3,714.8	0.54	0.67
Ocupats	2,777.2	28.4	2,721.4	2,832.9	1.02	1.30
Aturats	898.5	25.0	849.5	947.6	2.78	1.96
At. no treb.	84.2	8.9	66.7	101.6	10.60	2.29
At. sí treb.	814.4	23.0	769.3	859.5	2.83	1.80
Inactius	2,282.3	19.9	2,243.2	2,321.3	0.87	0.67

3.2. Estimació de la variància amb Balanced Repeated Replication.

El mètode de BRR és ideal quan hi ha dues unitats finals mostrals per clúster (o dos elements per estrat, en mostres simples), que no és el nostre cas.

Proposem una millora al mètode que aplica l'INE, sent més fidels al disseny de la mostra.

Recordem que per calcular la variabilitat de l'estadístic, l'INE divideix la mostra en dues meitats, incorporant aleatòriament la meitat de les seccions censals a la primera, i l'altra, a la segona. Aquest procés el reitera 20 cops, havent obtingut 40 meitats de la mostra i havent calculat la variància 40 cops, de les quals en treu una mitjana i calcula el coeficient de variació.

La nostra proposta no es basa en aquestes assignacions aleatòries, sinó en l'ús de les matrius de Hadamard, pretenent aconseguir la independència entre les mostres. Les matrius de Hadamard són matrius quadrades amb valors 1 o -1 i totes les files ortogonals entre elles. També hem revisat la modificació Fay, emprada sovint en la literatura, que és una generalització del mètode, i suposa un compromís entre el BRR tradicional i el Jackknife. Amb aquesta, en lloc d'utilitzar meitats iguals, utilitzem la mostra sencera però amb ponderacions diferents: k per les unitats fora de la meitat, i 2-k per les de dins. És també utilitzada per obtenir estadístics no lineals.

3.3. Estimació de la variància amb Jackknife.

En aquest cas, no utilitzem el Jackknife clàssic, sinó el Jackknife-n (per grups) [**44**], que és adient per quan hi ha més de 2 PSUs per estrat. Aquest consisteix en eliminar un grup per iteració, en lloc d'una única dada. Nosaltres eliminarem cada PSU.

3.4. Estimació de la variància amb Bootstrap.

Aquí tampoc utilitzem les versions clàssiques de Bootstrap, en ser la nostra una enquesta amb disseny complex. Fem servir el Bootstrap reescalat de Preston per a mostreigs complexos[**35**].

4. Estimació amb els factors de l'INE

Al web de l'INE, trobem publicades les magnituds de treball. Nosaltres les hem pogut reproduir fàcilment a partir dels factors facilitats, mitjançant les funcions surveysummary. Recordem, però, que per als càlculs del 2013 utilitzarem el calibratge antic, consultable encara al web de l'INE, que no té en compte la mida de la llar ni la nova variable de sexe i edat per província, i que utilitza com a base les dades del Cens de població i habitatges de 2001 en lloc del de 2011.

Presentem les taules publicades amb les magnituds i amb els coeficients de variació que han obtingut (utilitzant el seu mètode, amb 40 semimostres): Taula 6.4 i Taula 6.5.

Taula 6.4. Resultats del 2013 publicats al web de l'INE

2013	**T-1**	CV	**T-2**	CV	**T-3**	CV	**T-4**	CV
Actius	3,677.9	0.6	3,660.4	0.7	3,680.2	0.53	3,685.5	0.44
Ocupats	2,775.7	1.03	2,787.5	1.07	2,839.7	1.03	2,865.1	0.94
Aturats	902.3	3.05	873	2.58	840.5	2.85	820.4	2.6
At. no treb.	84.1	11.07	96.7	7.97	101	8.54	87.9	8.57
At. sí treb.	818.2	3	776.3	2.65	739.5	2.69	732.5	2.55
Inactius	2,280.1	0.97	2,291.8	1.13	2,236.4	0.88	2,280.6	0.72

Taula 6.5. Resultats del 2014 publicats al web de l'INE

2014	**T-1**	CV	**T-2**	CV	**T-3**	CV	**T-4**	CV
Actius	3,800.9	0.44	3,810.4	0.6	3,800.8	0.49	3,804.6	0.52
Ocupats	2,960.7	0.87	3,040.0	0.92	3,074.8	1.03	3,048.1	0.75
Aturats	840.2	2.4	770.4	3.07	726.1	3.79	756.5	3.18
At. no treb.	79.6	9.6	82.1	7.06	93.3	8.72	88	6.89
At. sí treb.	760.6	2.55	688.3	3.03	632.8	3.87	668.5	3.43
Inactius	2,275.5	0.74	2,267.5	1.02	2,269.2	0.83	2,272.1	0.87

Capítol 7
Discussió

A continuació, mostrem les estimacions de la població catalana de 16 i més anys en relació a l'activitat econòmica amb cadascun dels mètodes. Ho farem per al primer trimestre de 2013 i el tercer trimestre de 2014, ja que malgrat que són dos èpoques amb comportaments diferents en el mercat laboral i dues formes de calibrar diferents, degut a la millora del calibratge en 2014, ens resulten molts similars, com en els altres trimestres.

1. Comparació dels estimadors

A través de les estimacions dels totals per activitat econòmica, tenint com a referència les dades de l'INE, ens disposem a comparar els estimadors.

Taula 7.1. Comparació dels estimadors el primer trimestre de 2013

T1-2013	**H-T**	CV	**Raó S.**	CV	**Calibrat***	**INE**	CV
Actius	3,849.14	1.70	3,409.82	1.64	3,675.71	3,677.90	0.6
Ocupats	2,939.02	1.93	2,603.30	1.87	2,777.17	2,775.70	1.03
Aturats	910.12	3.65	806.52	3.49	898.55	902.3	3.05
At. no treb.	84.50	11.60	74.28	11.12	84.16	84.1	11.07
At. sí treb.	825.62	3.62	732.24	3.46	814.39	818.2	3
Inactius	2,881.65	1.85	2,548.18	1.79	2,282.28	2,280.10	0.97

Prenent les dades de l'INE com a fiables, que inclouen el tractament de la no-resposta, observem que l'estimador de Horvitz-Thompson sobreestima la població catalana de 16 anys o més. En canvi, l'estimador de raó separat no ho fa sobre el total, però sí que considera una població activa molt menor, en la qual la població d'ocupats és més petita però no en la mateixa magnitud que la població activa, i per tant la taxa d'atur resultarà més baixa (23.65% i 19.10% pels 24.53% i 19.23% de l'INE, durant el primer trimestre de 2013 i el tercer trimestre de 2014, respectivament, Taula 7.1 i Taula 7.2). Els coeficients de variació dels primers estimadors que hem calculat també són molt alts, i amb tot, els descartarem.

TAULA 7.2. Comparació dels estimadors el tercer trimestre de 2014

T3-2014	**H-T**	CV	**Raó S.**	CV	**Calibrat***	**INE**	CV
Actius	4,237.94	1.87	3,551.07	1.83	3805.55	3,800.80	0.49
Ocupats	3,421.00	2.05	2,868.10	1.98	3076.81	3,074.80	1.03
Aturats	816.94	3.65	682.96	3.56	728.74	726.1	3.79
At. no treb.	110.64	8.79	93.16	8.49	92.88	93.3	8.72
At. sí treb.	706.30	3.75	589.80	3.64	635.86	632.8	3.87
Inactius	3,016.34	2.01	2,518.99	1.92	2264.51	2,269.20	0.83

Si comparem els factors d'elevació, reforcem aquesta idea que els dels estimadors de Horvitz-Thompson i de raó separat tenen una distribució molt diferent als que l'INE facilita (FIGURA 7.1). En el cas de l'estimador calibrat -amb les dades d'entrada del de raó separat i aportant les variables auxiliars-, sembla ser que les ponderacions són molt similars a les oficials, i la correlació entre les dues és molt alta (TAULA 7.3).

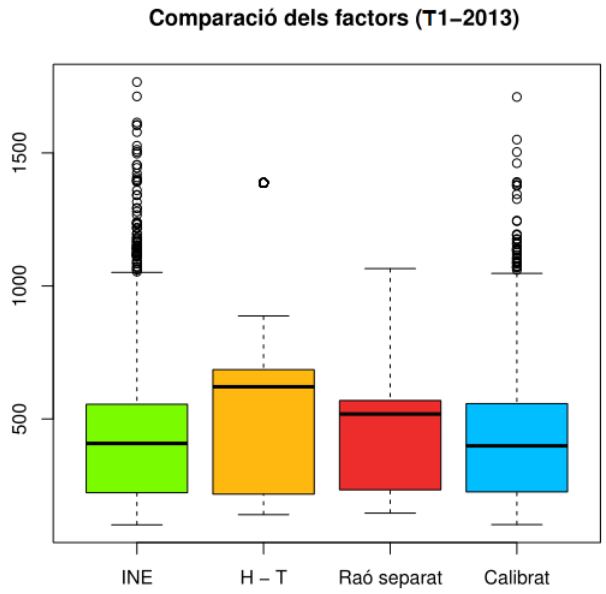

FIGURA 7.1. Boxplot per comparar els diferents factors

TAULA 7.3. Matriu de correlació dels factors

T1-2013	INE	H-T	Raó s.	Calibrat
INE	1.000	0.862	0.867	0.969
H-T	0.862	1.000	0.994	0.889
Raó s.	0.867	0.994	1.000	0.895
Calibrat	0.969	0.889	0.895	1.000

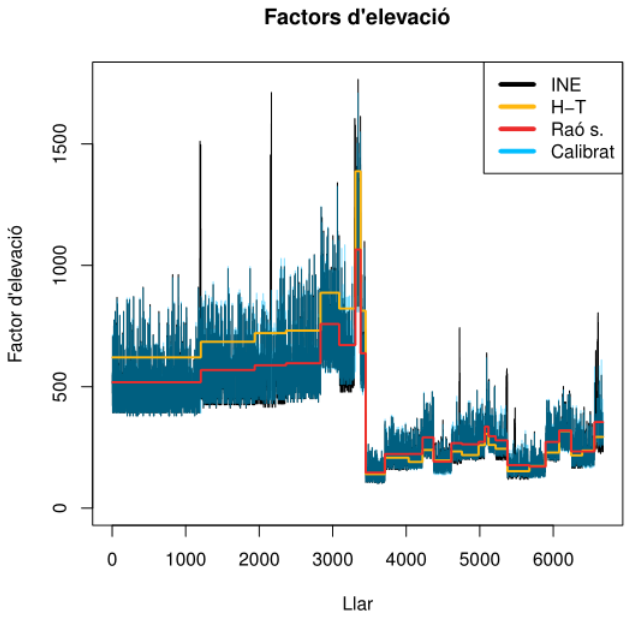

FIGURA 7.2. Factors d'elevació superposats
Hem aplicat un color blau cel amb transparència per al factor calibrat, que es veu fosc en superposició amb els factors de l'INE. Es visualitza la semblança entre aquests dos.

2. Validació de l'estimador calibrat per estimar els factors d'elevació

Per determinar si realment podem prendre les nostres dades per l'estudi, anem a comprovar si podem donar per vàlids els nostres factors calibrats, tot comparant-los amb els facilitat per l'INE, que tenen en compte la manca de resposta.

Després d'haver comprovat (ANNEX C) que les distribucions dels factors no segueixen una distribució normal, ens hem proposat realitzar un test no paramètric: la prova de Kolmogorov-Smirnov. Aquesta prova ens dóna un estadístic D, la distància de Smirnov, una distància entre la funció de distribució empírica dels nostres pesos i la funció de distribució dels de l'INE: la màxima diferència, en valor absolut, de les dues corbes.
Volem testar:

$$\begin{cases} H_0 : Els\ factors\ calibrats\ es\ distribueixen\ com\ els\ de\ l'INE \\ H_A : \neg H_0 \end{cases}$$

L'executem mitjançant la instrucció ks.test d'R.

```
        Two-sample Kolmogorov-Smirnov test

data:  factors$INE and factors$calib
D = 0.027, p-value = 0.01563
alternative hypothesis: two-sided
```

Pel resultat obtingut, podem dir que hi ha diferències significatives. Però si ens fixem bé, ens dóna una distància de Smirnov de 0.027! A més, tenint una mostra tan gran (de més de 6000 llars), és molt fàcil que es rebutgi la hipòtesi nul·la en un test[1]. Si representem les distribucions mencionades abans, podem veure com es tracta d'una distància molt petita FIGURA 7.3, que ens sembla acceptable.

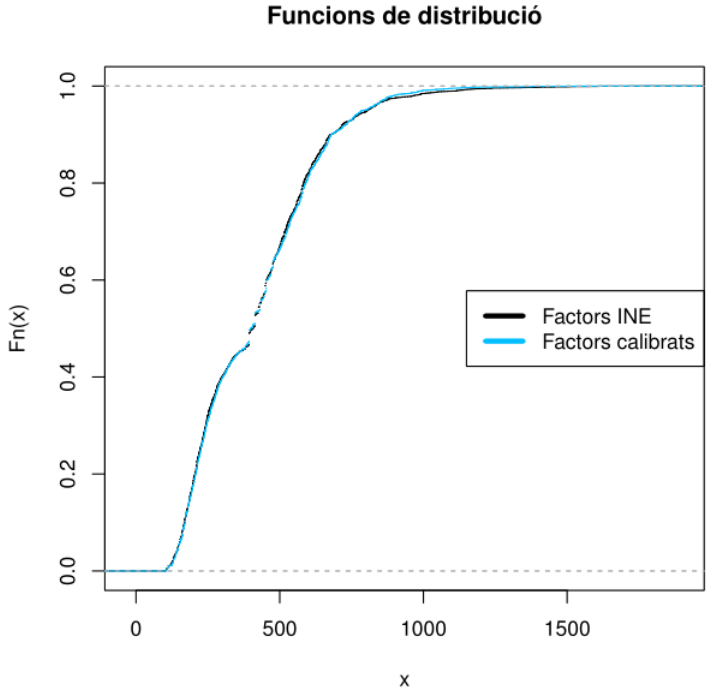

FIGURA 7.3. Funció de distribució empírica dels factors

Finalment, afirmem que l'estimador calibrat a partir de l'estimador de raó separat ens dóna uns valors molt semblants als de l'INE i els prenem per a seguir l'estudi. A partir d'ara, ens centrarem en els resultats dels errors de l'estimador calibrat utilitzant pesos replicats (tècniques de remostreig) i la linealització de Taylor.

3. Comparació dels estimadors calibrats

Per trobar els valors veritables dels totals poblacionals i taxes, necessitaríem fer un cens en un temps molt reduït (i comprovar la veracitat de les dades). En el cas

[1]Ho apunta Yates i ho recull Kish a: Kish, L., Muestreo de Encuestas, Trillas 1972, pàg. 678.

dels errors mostrals, nosaltres no podrem conèixer la realitat, ni podrem conèixer amb seguretat quin és el millor mètode. Tenim, però, diverses qualitats que ens poden ajudar a triar-ne un: que s'executi ràpidament, que sigui constant en cada execució i en el temps, que generi un nombre mínim de rèpliques per obtenir una "bona" aproximació, que sigui senzill (i possible) d'implementar i d'entendre, que estigui integrat a la majoria del programari estadístic, que ens permeti d'obtenir d'altres tipus d'estadístics, com ho són els no lineals... Amb aquesta idea global, volem comparar els diferents mètodes proposats, amb estimador calibrat: Taylor, BRR, Jackknife i Bootstrap, amb les eines que disposem.

3.1. Cost computacional.

Com que tots els estimadors s'han obtingut en un temps i amb uns recursos raonables, no creiem convenient que el cost computacional sigui un motiu per descartar-ne cap, almenys per una població com la de Catalunya, a no ser que s'utilitzin moltes rèpliques de Bootstrap. A la FIGURA 7.4, podem veure representat el temps mitjà que ha tardat el procés de calibratge (amb les replicacions) amb cada mètode, i fent 50 execucions. Observem que la versió utilitzada de Jackknife és més de 10 cops més lent que la linealització. Els Bootstrap-50, Bootstrap-100 i Bootstrap-200 són bootstraps MRB amb 50 i 200 replicacions, respectivament. La linealització de Taylor, en no necessitar pesos replicats, s'ha executat molt ràpidament.

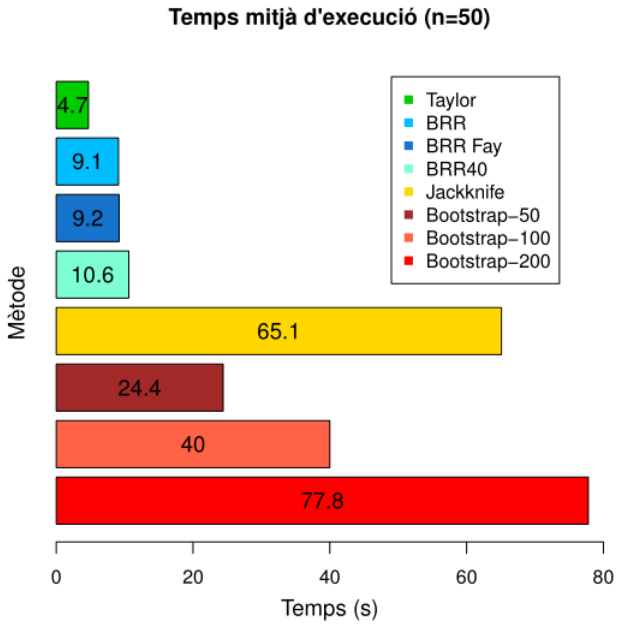

FIGURA 7.4. Comparació del temps mitjà d'execució dels diferents mètodes

Característiques de l'ordinador.

Tots els càlculs s'han realitzat amb un ordinador amb les característiques següents:

- Processador Intel Core i5-3317U @ 1.70GHz

- Sistema operatiu de 64 bits
- Memòria RAM de 6GB
- Windows 8.1
- Versió d'RStudio: 0.98.1103
- Versió d'R: 3.1.0

Cal destacar que el cas del Bootstrap hem intentat trobar una n òptima per la qual els valors s'estabilitzin cada cop que executem l'algoritme, ja que coneixem que ho fa a mesura que augmenta[**3**]. Fins i tot amb un algoritme iteratiu on vam definir una tolerància mínima acceptable de variabilitat, havíem d'aturar la seva execució després de moltes hores, així que creiem que aquest mètode requereix una potència computacional molt alta. Effron i Tibshirani (1994)[**15**] recomanaven una n entre 50 i 200.

3.2. Tria entre els estimadors amb BRR.

A la Taula 7.4 i a la Taula 7.5, de coeficients de variació en percentatge, rho indica el paràmetre de la modificació de Fay, on $\rho = 0$ és el mètode BRR tradicional, i H 40 fa referència a una matriu Hadamard de dimensió 40 per comparar amb les dimensions de semimostres de l'INE (quan per defecte, n'utilitzem menys). Per valors de rho superiors a 1, no hem obtingut estimacions (és indiferent donar-li un pes 0.5 a una meitat i un 1.5 a l'altra, que fer-ho a l'inrevés). Com podem veure, els coeficients de variació són molt semblants en línies generals, però sí que sembla ser que el fet d'utilitzar una matriu Hadamard de més dimensió no ens és d'ajuda (de fet, la matriu de Hadamard tindrà dimensió n x n, on n serà el múltiple de 4 que excedeixi menys del nombre de factors que hi ha al calibratge). Mostrem dos trimestres perquè les petites disminucions dels coeficients que hi ha en uns trimestres per a certs rho, no hi són en altres. En conclusió, podríem utilitzar el primer BRR, sense cap modificació.

3.3. Tria del mètode.

Els estimadors calibrats presenten uns valors propers a les dades facilitades per l'INE (biaix menor) amb uns errors molt més baixos. A la Figura3.3, observem el següent:

- El Jackknife n i la linealització de Taylor són molt semblants, com ja havia citat Berger (2008)[**8**].
- A més, el Jackknife utilitzat i la linealització de Taylor presenten corbes suaus, és a dir, la seva variació és bastant constant en el temps.
- A cada execució de l'algoritme que hem fet, les estimacions de cada mètode han estat idèntiques. Aquest no és el cas del Bootstrap, però, que ha anat variant molt.
- El Bootstrap amb 100 replicacions, que sembla funcionar bastant bé aquí, podria haver presentat valors bastant diferents, com es pot intuir a partir de les Taules 7.6 i 7.7. El desestimem.

Ens decantaríem, doncs, per utilitzar el mètode de la linealització.

Per últim, si mirem la Figura 7.5, veiem que no podríem notar gaire diferències entre els estimadors, ni tan sols amb els de l'INE.

TAULA 7.4. Coeficients de variació de diferents BRR per estimar els totals per activitat econòmica (T1-2013)

T1-2013	**Actius**	**Ocupats**	**Aturats**	**At. no treb.**	**At. sí treb.**	**Inactius**
Total	3,675.71	2,777.17	898.55	84.16	814.39	2,282.29
ρ=**0**	0.4985	0.9394	2.2706	10.1079	2.1059	0.8029
ρ=**0.1**	0.4979	0.9389	2.2685	10.0998	2.1028	0.8020
ρ=**0.2**	0.4974	0.9386	2.2674	10.0948	2.1004	0.8011
ρ=**0.3**	0.4970	0.9385	2.2671	10.0927	2.0989	0.8004
ρ=**0.4**	0.4967	0.9386	2.2678	10.0936	2.0981	0.7999
ρ=**0.5**	0.4964	0.9390	2.2694	10.0975	2.0981	0.7995
ρ=**0.6**	0.4962	0.9396	2.2719	10.1044	2.0989	0.7992
ρ=**0.7**	0.4961	0.9404	2.2753	10.1142	2.1004	0.7990
ρ=**0.8**	0.4961	0.9414	2.2796	10.1271	2.1026	0.7990
ρ=**0.9**	0.4962	0.9427	2.2848	10.1431	2.1057	0.7991
ρ=**0.95**	0.4962	0.9434	2.2877	10.1522	2.1075	0.7992
ρ=**0.99**	0.4963	0.9441	2.2902	10.1601	2.1090	0.7993
H 40	0.5000	0.9679	2.3801	10.2099	2.2075	0.8053

TAULA 7.5. Coeficients de variació de diferents BRR per estimar els totals per activitat econòmica (T3-2014)

T3-2014	**Actius**	**Ocupats**	**Aturats**	**At. no treb.**	**At. sí treb.**	**Inactius**
Total	3,805.55	3,076.81	728.74	92.88	635.86	2,264.51
ρ=**0**	0.5684	0.5912	2.5255	7.1292	3.1289	0.9552
ρ=**0.1**	0.5664	0.5897	2.5216	7.0958	3.1236	0.9518
ρ=**0.2**	0.5646	0.5885	2.5185	7.0665	3.1192	0.9488
ρ=**0.3**	0.5630	0.5876	2.5162	7.0414	3.1157	0.9461
ρ=**0.4**	0.5615	0.5870	2.5147	7.0201	3.1131	0.9437
ρ=**0.5**	0.5603	0.5867	2.5140	7.0026	3.1113	0.9416
ρ=**0.6**	0.5592	0.5866	2.5141	6.9887	3.1103	0.9398
ρ=**0.7**	0.5583	0.5868	2.5150	6.9784	3.1102	0.9383
ρ=**0.8**	0.5576	0.5873	2.5166	6.9716	3.1109	0.9371
ρ=**0.9**	0.5571	0.5880	2.5190	6.9682	3.1123	0.9361
ρ=**0.95**	0.5568	0.5884	2.5205	6.9678	3.1134	0.9358
ρ=**0.99**	0.5567	0.5888	2.5219	6.9681	3.1144	0.9355
H 40	0.5659	0.5912	2.5709	7.1423	3.1714	0.9511

TAULA 7.6. Coeficients de variació i variabilitat d'aquests, per les diferents tècniques, el primer trimestre de 2013

T1-2013	Taylor	SD	Bootstrap	SD	Jackknife	SD	BRR	SD
Actius	0.5422	0	0.5472	0.0316	0.5455	0	0.4985	0
Ocupats	1.0242	0	1.0255	0.0841	1.0307	0	0.9394	0
Aturats	2.7847	0	2.7143	0.1519	2.8040	0	2.2706	0
At. no treb.	10.5985	0	10.7612	0.5850	10.7513	0	10.1079	0
At. sí treb.	2.8274	0	2.7489	0.1606	2.8432	0	2.1059	0
Inactius	0.8733	0	0.8813	0.0509	0.8786	0	0.8029	0

TAULA 7.7. Coeficients de variació i variabilitat d'aquests, per les diferents tècniques, el tercer trimestre de 2014

T3-2014	Taylor	SD	Bootstrap	SD	Jackknife	SD	BRR	SD
Actius	0.4957	0	0.5088	0.0418	0.4991	0	0.5684	0
Ocupats	0.8575	0	0.8695	0.0559	0.8644	0	0.5912	0
Aturats	3.0381	0	3.0018	0.2337	3.0642	0	2.5255	0
At. no treb.	7.9596	0	8.2534	0.5459	8.0475	0	7.1292	0
At. sí treb.	3.1948	0	3.1492	0.2231	3.2190	0	3.1289	0
Inactius	0.8330	0	0.8551	0.0703	0.8388	0	0.9552	0

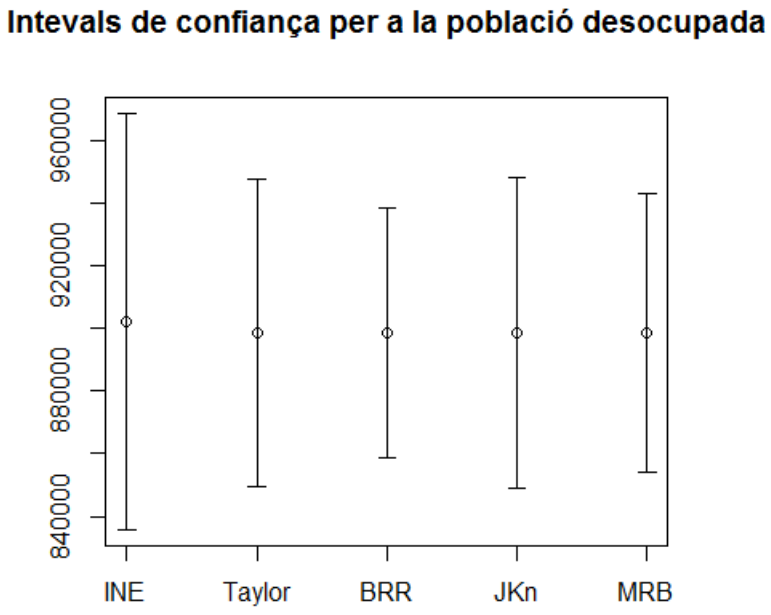

FIGURA 7.5. ICs dels diferents mètodes per a l'estimació de la població desocupada

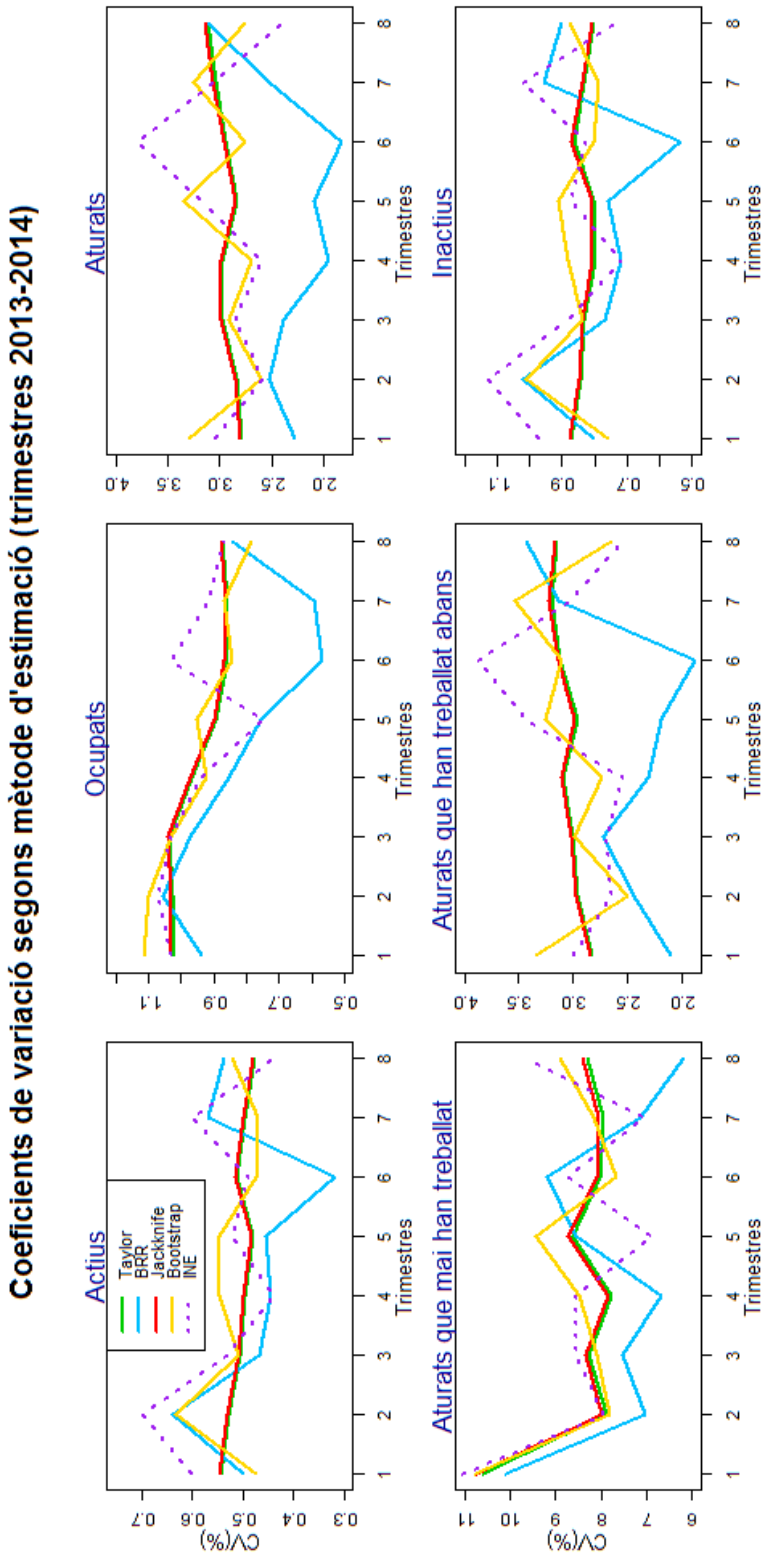

FIGURA 7.6. Coeficients de variació dels diferents mètodes d'estimació durant els trimestres de 2013 i 2014
Aquests coeficients s'han obtingut en una única execució (només el Bootstrap variarà, si l'executem més cops).

Capítol 8
Resultats

L'objectiu del treball era obtenir estimacions de l'error per les principals magnituds de treball. Aquí en mostrarem les del primer trimestre de 2013 amb la metodologia que hem triat: la linealització de Taylor. Notem com augmenten els coeficients de variació en els col·lectius més minoritaris.

1. Població activa, ocupada i desocupada

Per calcular els totals poblacionals, hem fet servir la funció svytotal. Recordem que venen expressats en milers.

1.1. Població activa.

Taula 8.1. Població activa (T1-2013)

	Total	SE	IC 2.5%	IC 97.5%	CV(%)
Població activa	3,675.7	19.9	3,636.6	3,714.8	0.54

Taula 8.2. Població activa per sexe (T1-2013)

	Total	SE	IC 2.5%	IC 97.5%	CV(%)
Home	1,947.2	10.5	1,926.7	1,967.6	0.54
Dona	1,728.6	16.1	1,697.1	1,760.1	0.93

Taula 8.3. Població activa per grup d'edat (T1-2013)

	Total	SE	IC 2.5%	IC 97.5%	CV(%)
16-19	53.9	4.7	44.7	63.1	8.70
20-24	231.6	6.6	218.7	244.5	2.84
25-54	2,881.2	13.7	2,854.3	2,908.1	0.48
55i+	509.1	10.8	487.9	530.3	2.13

Taula 8.4. Població activa per nacionalitat (T1-2013)

	Total	SE	IC 2.5%	IC 97.5%	CV(%)
Espanyola	3,060.8	17.8	3,025.9	3,095.6	0.58
No espanyola	614.9	9.5	596.3	633.6	1.54

Taula 8.5. Població activa per província (T1-2013)

	Total	SE	IC 2.5%	IC 97.5%	CV(%)
Barcelona	2,673.5	20.1	2,634.1	2,712.8	0.75
Girona	375.3	6.8	361.8	388.7	1.83
Lleida	208.5	7.7	193.4	223.6	3.70
Tarragona	418.5	8.2	402.5	434.5	1.95

Taula 8.6. Població activa per nivell de formació assolit (T1-2013)

	Total	SE	IC 2.5%	IC 97.5%	CV(%)
Analfabets i educació primària	520.8	24.4	472.9	568.7	4.69
Educació secundària 1a. etapa	989.9	30.6	929.9	1,049.9	3.09
Educació secundària 2a. etapa	838.9	24.4	791.0	886.8	2.91
Educació superior	1,326.1	37.1	1,253.3	1,398.9	2.80

Taula 8.7. Població activa per sector econòmic (T1-2013)

	Total	SE	IC 2.5%	IC 97.5%	CV(%)
Agricultura	58.9	6.8	45.5	72.2	11.55
Indústria	583.5	20.3	543.7	623.3	3.48
Construcció	214.0	11.1	192.3	235.7	5.17
Serveis	2,288.4	32.0	2,225.7	2,351.1	1.40

TAULA 8.8. Població activa per branca d'activitat (T1-2013)

	Total	SE	IC 2.5%	IC 97.5%	CV(%)
Agricultura, ramaderia, silvicultura i pesca	58.9	6.8	45.5	72.2	11.55
Indústries extractives, energia, aigua i residus	48.0	5.6	37.1	58.9	11.57
Alimentació, tèxtil, fusta, paper i arts gràfiques	188.7	13.8	161.8	215.7	7.29
Química i cautxú	89.9	7.3	75.6	104.1	8.10
Metal·lúrgia	74.6	6.4	62.0	87.2	8.59
Maquinària, material elèctric i de transport	182.4	12.8	157.4	207.3	6.99
Construcció	214.0	11.1	192.3	235.7	5.17
Comerç engròs i reparació vehicles motor i motocicletes	167.4	10.3	147.2	187.6	6.16
Comerç detall	322.4	15.7	291.7	353.1	4.85
Transport i emmagatzematge	178.6	10.0	159.0	198.1	5.59
Hostaleria	226.0	13.4	199.8	252.2	5.92
Informació i comunicacions	98.4	8.4	81.9	114.8	8.54
Activitats financeres i assegurances	79.2	6.7	66.0	92.4	8.50
Activitats immobiliàries, professionals i tècniques	201.2	13.3	175.1	227.2	6.61
Activitats administratives i serveis auxiliars	170.9	9.8	151.6	190.2	5.76
Administració pública	143.2	9.9	123.8	162.6	6.91
Educació	189.3	11.5	166.8	211.8	6.07
Sanitat i serveis socials	260.9	12.9	235.6	286.1	4.93

1.2. Població ocupada.

Taula 8.9. Població ocupada (T1-2013)

	Total	SE	IC 2.5%	IC 97.5%	CV(%)
Població ocupada	2,777.2	28.4	2,721.4	2,832.9	1.02

Taula 8.10. Població ocupada per sexe (T1-2013)

	Total	SE	IC 2.5%	IC 97.5%	CV(%)
Home	1,457.6	16.9	1,424.5	1,490.6	1.16
Dona	1,319.6	19.2	1,281.9	1,357.3	1.46

Taula 8.11. Població ocupada per grup d'edat (T1-2013)

	Total	SE	IC 2.5%	IC 97.5%	CV(%)
16-19	12.5	2.3	8.0	17.0	18.43
20-24	124.4	7.3	110.0	138.7	5.91
25-54	2,225.4	22.5	2,181.4	2,269.4	1.01
55i+	415.0	11.5	392.4	437.5	2.77

Taula 8.12. Població ocupada per nacionalitat (T1-2013)

	Total	SE	IC 2.5%	IC 97.5%	CV(%)
Espanyola	2,419.9	25.3	2,370.4	2,469.4	1.04
No espanyola	357.3	12.9	331.9	382.6	3.62

Taula 8.13. Població ocupada per província (T1-2013)

	Total	SE	IC 2.5%	IC 97.5%	CV(%)
Barcelona	2,025.4	27.1	1,972.3	2,078.5	1.34
Girona	280.2	6.8	267.0	293.5	2.41
Lleida	173.5	7.8	158.2	188.8	4.50
Tarragona	298.0	8.8	280.7	315.3	2.97

Taula 8.14. Població ocupada per nivell de formació assolit (T1-2013)

	Total	SE	IC 2.5%	IC 97.5%	CV(%)
Analfabets i educació primària	315.4	15.6	284.7	346.1	4.96
Educació secundària 1a. etapa	679.3	21.1	637.8	720.7	3.11
Educació secundària 2a. etapa	647.9	21.6	605.5	690.3	3.34
Educació superior	1,134.6	34.0	1,067.9	1,201.3	3.00

Taula 8.15. Població ocupada per sector econòmic (T1-2013)

	Total	SE	IC 2.5%	IC 97.5%	CV(%)
Agricultura	50.8	6.4	38.4	63.3	12.52
Indústria	521.3	18.7	484.6	558.0	3.59
Construcció	173.2	10.7	152.1	194.2	6.21
Serveis	2,031.9	31.1	1,971.0	2,092.8	1.53

TAULA 8.16. Població ocupada per branca d'activitat (T1-2013)

	Total	SE	IC 2.5%	IC 97.5%	CV(%)
Agricultura, ramaderia, silvicultura i pesca	50.8	6.4	38.4	63.3	12.52
Indústries extractives, energia, aigua i residus	42.6	4.9	33.0	52.2	11.55
Alimentació, tèxtil, fusta, paper i arts gràfiques	166.9	12.6	142.2	191.5	7.55
Química i cautxú	81.8	6.9	68.2	95.4	8.48
Metal·lúrgia	66.5	6.2	54.3	78.7	9.39
Maquinària, material elèctric i de transport	163.5	11.8	140.3	186.7	7.24
Construcció	173.2	10.7	152.1	194.2	6.21
Comerç engròs i reparació vehicles motor i motocicletes	152.7	9.8	133.5	172.0	6.42
Comerç detall	279.7	14.7	250.9	308.5	5.26
Transport i emmagatzematge	153.9	9.8	134.8	173.1	6.36
Hostaleria	178.9	12.0	155.4	202.5	6.71
Informació i comunicacions	89.3	7.4	74.7	103.9	8.34
Activitats financeres i assegurances	74.7	6.6	61.8	87.7	8.85
Activitats immobiliàries, professionals i tècniques	178.6	12.4	154.3	203.0	6.95
Activitats administratives i serveis auxiliars	145.1	9.1	127.3	162.8	6.24
Administració pública	136.4	9.5	117.7	155.1	7.00
Educació	181.9	11.3	159.8	204.1	6.21
Sanitat i serveis socials	241.9	12.7	217.1	266.7	5.23
Activitats culturals i esportives, i altres serveis	218.6	12.1	195.0	242.3	5.51

1.3. Població desocupada.

Taula 8.17. Població desocupada (T1-2013)

	Total	SE	IC 2.5%	IC 97.5%	CV(%)
Població desocupada	898.5	25.0	849.5	947.6	2.78

Taula 8.18. Població desocupada per sexe (T1-2013)

	Total	SE	IC 2.5%	IC 97.5%	CV(%)
Home	489.6	16.2	457.9	521.2	3.30
Dona	409.0	15.9	377.8	440.1	3.89

Taula 8.19. Població desocupada per grup d'edat (T1-2013)

	Total	SE	IC 2.5%	IC 97.5%	CV(%)
16-19	41.4	4.3	33.0	49.8	10.33
20-24	107.2	6.6	94.2	120.2	6.19
25-54	655.8	19.8	617.0	694.5	3.02
55i+	94.1	7.2	79.9	108.3	7.69

Taula 8.20. Població desocupada per nacionalitat (T1-2013)

	Total	SE	IC 2.5%	IC 97.5%	CV(%)
Espanyola	640.9	21.0	599.8	682.0	3.27
No espanyola	257.7	14.3	229.6	285.8	5.56

Taula 8.21. Població desocupada per província (T1-2013)

	Total	SE	IC 2.5%	IC 97.5%	CV(%)
Barcelona	648.1	24.0	601.0	695.2	3.71
Girona	95.0	6.2	82.9	107.1	6.49
Lleida	35.0	4.5	26.1	43.9	12.96
Tarragona	120.5	8.5	103.9	137.0	7.02

Taula 8.22. Població desocupada per nivell de formació assolit (T1-2013)

	Total	SE	IC 2.5%	IC 97.5%	CV(%)
Analfabets i educació primària	205.4	15.9	174.2	236.6	7.75
Educació secundària 1a. etapa	310.6	17.8	275.8	345.5	5.73
Educació secundària 2a. etapa	191.0	10.1	171.2	210.8	5.29
Educació superior	191.5	11.6	168.9	214.2	6.04

Taula 8.23. Població desocupada per sector econòmic (T1-2013)

Aquesta taula segurament no es publicaria degut a que el sector econòmic és aquell en el qual la persona s'hi dedica o s'hi ha dedicat fins un any abans. Amb tants aturats de llarga durada, aquesta vinculació perd sentit.

	Total	SE	IC 2.5%	IC 97.5%	CV(%)
Agricultura	8.0	1.8	4.4	11.6	22.77
Indústria	62.2	6.4	49.7	74.8	10.31
Construcció	40.8	5.1	30.9	50.8	12.42
Serveis	256.5	12.9	231.1	281.8	5.04

2. Taxes d'activitat, d'ocupació i d'atur

Per calcular les taxes, hem fet servir la funció svymean i hem multiplicat els resultats per 100.

2.1. Taxes d'activitat.

Taula 8.24. Taxa d'activitat (T1-2013)

	Taxa(%)	SE(%)	IC 2.5%	IC 97.5%	CV(%)
Taxa d'activitat	61.69	0.33	61.04	62.35	0.54

Taula 8.25. Taxa d'activitat per sexe (T1-2013)

	Taxa(%)	SE(%)	IC 2.5%	IC 97.5%	CV(%)
Home	67.85	0.36	67.14	68.56	0.54
Dona	55.97	0.52	54.95	56.99	0.93

TAULA 8.26. Taxa d'activitat per grup d'edat (T1-2013)

	Taxa(%)	SE(%)	IC 2.5%	IC 97.5%	CV(%)
16-19	20.38	1.77	16.91	23.86	8.70
20-24	66.30	1.88	62.61	69.99	2.84
25-54	89.56	0.43	88.72	90.39	0.48
55i+	23.93	0.51	22.94	24.93	2.13

TAULA 8.27. Taxa d'activitat per nacionalitat (T1-2013)

	Taxa(%)	SE(%)	IC 2.5%	IC 97.5%	CV(%)
Espanyola	59.30	0.34	58.63	59.98	0.58
No espanyola	77.20	1.19	74.86	79.54	1.54

TAULA 8.28. Taxa d'activitat per província (T1-2013)

	Taxa(%)	SE(%)	IC 2.5%	IC 97.5%	CV(%)
Barcelona	61.48	0.46	60.58	62.39	0.75
Girona	63.08	1.15	60.83	65.34	1.83
Lleida	58.38	2.16	54.14	62.61	3.70
Tarragona	63.63	1.24	61.20	66.06	1.95

TAULA 8.29. Taxa d'activitat per nivell de formació assolit (T1-2013)

	Taxa(%)	SE(%)	IC 2.5%	IC 97.5%	CV(%)
Analfabets i educació primària	28.46	0.91	26.67	30.25	3.21
Educació secundària 1a. etapa	71.28	1.16	68.99	73.56	1.63
Educació secundària 2a. etapa	70.67	1.00	68.72	72.62	1.41
Educació superior	85.44	0.73	84.01	86.86	0.85

2.2. Taxes d'ocupació.

Taula 8.30. Taxa d'ocupació (T1-2013)

	Taxa(%)	SE(%)	IC 2.5%	IC 97.5%	CV(%)
Taxa d'ocupacio	46.61	0.48	45.68	47.55	1.02

Taula 8.31. Taxa d'ocupació per sexe (T1-2013)

	Taxa(%)	SE(%)	IC 2.5%	IC 97.5%	CV(%)
Home	50.79	0.59	49.64	51.94	1.16
Dona	42.73	0.62	41.51	43.95	1.46

Taula 8.32. Taxa d'ocupació per grup d'edat (T1-2013)

	Taxa(%)	SE(%)	IC 2.5%	IC 97.5%	CV(%)
16-19	4.71	0.87	3.01	6.41	18.43
20-24	35.60	2.10	31.48	39.72	5.91
25-54	69.17	0.70	67.80	70.54	1.01
55i+	19.51	0.54	18.45	20.57	2.77

Taula 8.33. Taxa d'ocupació per nacionalitat (T1-2013)

	Taxa(%)	SE(%)	IC 2.5%	IC 97.5%	CV(%)
Espanyola	46.88	0.49	45.92	47.84	1.04
No espanyola	44.85	1.62	41.67	48.03	3.62

Taula 8.34. Taxa d'ocupació per província (T1-2013)

	Taxa(%)	SE(%)	IC 2.5%	IC 97.5%	CV(%)
Barcelona	46.58	0.62	45.36	47.80	1.34
Girona	47.11	1.14	44.89	49.34	2.41
Lleida	48.58	2.19	44.29	52.86	4.50
Tarragona	45.31	1.34	42.68	47.95	2.97

TAULA 8.35. Taxa d'ocupació per nivell de formació assolit (T1-2013)

	Taxa(%)	SE(%)	IC 2.5%	IC 97.5%	CV(%)
Analfabets i educació primària	17.24	0.68	15.91	18.57	3.93
Educació secundària 1a. etapa	48.91	0.97	47.01	50.81	1.98
Educació secundària 2a. etapa	54.58	1.05	52.52	56.64	1.93
Educació superior	73.10	0.87	71.40	74.80	1.19

2.3. Taxes d'atur.

Taula 8.36. Taxa d'atur (T1-2013)

	Taxa(%)	SE(%)	IC 2.5%	IC 97.5%	CV(%)
Taxa d'atur	24.45	0.66	23.14	25.75	2.72

Taula 8.37. Taxa d'atur per sexe (T1-2013)

	Taxa(%)	SE(%)	IC 2.5%	IC 97.5%	CV(%)
Home	25.14	0.81	23.56	26.72	3.21
Dona	23.66	0.88	21.93	25.39	3.74

Taula 8.38. Taxa d'atur per grup d'edat (T1-2013)

	Taxa(%)	SE(%)	IC 2.5%	IC 97.5%	CV(%)
16-19	76.88	3.88	69.27	84.50	5.05
20-24	46.30	2.66	41.09	51.51	5.74
25-54	22.76	0.68	21.43	24.09	2.99
55i+	18.49	1.38	15.78	21.20	7.48

Taula 8.39. Taxa d'atur per nacionalitat (T1-2013)

	Taxa(%)	SE(%)	IC 2.5%	IC 97.5%	CV(%)
Espanyola	20.94	0.68	19.61	22.26	3.23
No espanyola	41.90	2.10	37.78	46.02	5.02

Taula 8.40. Taxa d'atur per província (T1-2013)

	Taxa(%)	SE(%)	IC 2.5%	IC 97.5%	CV(%)
Barcelona	24.24	0.87	22.54	25.95	3.59
Girona	25.32	1.48	22.41	28.23	5.87
Lleida	16.79	2.09	12.68	20.89	12.47
Tarragona	28.79	1.85	25.17	32.40	6.41

TAULA 8.41. Taxa d'atur per nivell de formació assolit (T1-2013)

	Taxa(%)	SE(%)	IC 2.5%	IC 97.5%	CV(%)
Analfabets i educació primària	39.44	1.99	35.54	43.34	5.05
Educació secundària 1a. etapa	31.38	1.26	28.91	33.85	4.01
Educació secundària 2a. etapa	22.77	1.07	20.67	24.86	4.69
Educació superior	14.44	0.80	12.88	16.00	5.51

Capítol 9
Conclusions

1. Contribucions i resultats principals

Fa dues dècades, s'havia fet un estudi semblant de Catalunya[**18**]. Ara, disposem de més dades per part de l'INE, de més facilitats computacionals i de més modificacions dels mètodes de remostreig.

Treballadors de l'INE també van fer un document de treball[**4**] per tal d'estudiar els demés mètodes, sent conscients que el de les semimostres reiterades que utilitza no és el més adient.

En aquest treball, hem fet una comparació de les tècniques per estimar l'error mostral, utilitzant versions revisades de les més clàssiques:

- Les mostres equilibrades, les hem utilitzat amb una matriu ortogonal, per tal d'abandonar l'evident dependència entre les mostres que hi havia quan partíem les PSUs en dos, ja que una semimostra contenia les dades que no tenia l'altra. També hem provat la modificació de Fay, però no ens ha aportat res. En general, aquest mètode ens ha funcionat bé i el calibratge s'executa en pocs segons.
- El Jackknife-n ha reemplaçat la versió clàssica del mètode, per tal de ser útil per al nostre disseny bietàpic.
- El Bootstrap utilitzat ha estat el reescalat de Preston per a mostreigs complexos, com era el nostre cas. Un problema que ha tingut és la variabilitat que presenta cada cop que s'executa, i que per evitar-ho, s'ha de recórrer a un ordinador potent per tal d'incrementar molt el nombre de reiteracions, amb la qual cosa el cost computacional és molt elevat.

Som conscients que no podrem saber els valors reals, però hem tingut en compte aspectes com el temps d'execució, la facilitat d'obtenció dels resultats, la interpretació dels mètodes o la implementació en el programa R a l'hora d'escollir un mètode per presentar els resultats del mercat de treball per Catalunya: la linealització de Taylor, que ens ha donat estabilitat a l'hora d'estimar l'error al llarg del temps, és el més ràpid en executar-se i ve recolzat per les propietats del Jackknife, que recordem que s'adapta fàcilment als diferents dissenys i tracta la no-resposta d'una manera senzilla. Segons les circumstàncies, ens pot anar bé comptar també amb

aquest últim: la linealització de Taylor i el Jackknife-n són mètodes enfocats d'una manera molt diferent.

A banda d'aquests resultats facilitats, podríem reproduir la metodologia en altres enquestes. I millor encara si són enquestes pròpies de l'Idescat i es pot conèixer l'efecte de la no-resposta.

Hem comprovat també la importància del calibratge en estimar totals i taxes sense obtenir CVs diferents.

Cal afegir que gràcies a les replicacions, podem permetre a un investigador obtenir una bona aproximació de l'error estàndard sense saltar-nos el secret estadístic. Com és lògic, aquesta persona necessitaria disposar de dades com podrien ser les seccions censals i els estrats, que no se'ns és permès de donar. En canvi sí que podria disposar d'una generació de pesos replicats.

Per últim, hem comprovat que l'R pot ser una eina fiable, i que pot ser una bona alternativa, entre d'altres coses, pel seu cost nul.

2. Limitacions i futura recerca

Hem partit d'unes dades que no incloïen els torns de rotació ni l'efecte de la manca de resposta. Hagués estat interessant poder estimar aquesta darrera per obtenir més precisió, cosa que podríem estudiar en un futur.

No hem tingut en compte la correcció per població finita, però és negligible.

Com hem comentat abans, una limitació molt gran és el fet de no poder fer censos per conèixer la realitat, que fa que no puguem saber mai si el mètode triat és el millor, tot i que tenim arguments suficients com per adoptar-ne algun, entrant en joc el pensament estadístic.

Seria bo provar les diferents funcions de calibratge (logit i raking) i afinar-ne els límits. El nostre treball s'ha basat en estimar l'error estàndard de l'INE, procurant apropar-nos a les taxes i als totals publicats, objectiu que s'ha complert.

D'altra banda, un estudi més profund dels mètodes (inclosa l'estimació amb fórmules teòriques), ajudaria a entendre millor les diferències observades en els factors obtinguts.

3. Valoració personal

Aquest treball ha comportat molt d'esforç, molts errors i molt d'aprenentatge acadèmic i personal. Ha suposat enfocar d'una altra manera una feina que, en principi, s'havia basat en la programació de mètodes de remostreig, fins que es va decidir explotar les capacitats del paquet *survey*, i ens vam poder centrar en temes més importants com ho és el calibratge.

Només dir que el viatge ha estat del tot interessant, ja que no només ha servit per intimar amb l'Enquesta de població activa, aprendre mètodes de remostreig, investigar sobre noves funcions en R i aconseguir processos més eficients, sinó que també m'ha permès veure el funcionament per dins de l'Institut d'Estadística de Catalunya i, fins i tot, interaccionar amb l'INE per resoldre alguns dels dubtes sorgits en el procés.

Bibliografia

[1] Adeshiyan, S. A.; Thompson, K. J. & Ozcoskun, L. T. (2007), Investigation of Jackknife Linearization Variance Estimators for US Census Bureau Business Survey Estimates, in 'Proceedings of the Third International Conference on Business Surveys'.

[2] Ardilly, P. & Osier, G. (2007), Cross-sectional variance estimation for the French "Labour Force Survey.", in 'Survey Research Methods', pp. 75–83.

[3] Axelson, M.; Di Consiglio, L.; Djerf, K.; Falorsi, S.; Kowarik, A.; Liberts, M.; Nikolaidis, I.; Berger, Y.; Münnich, R.; Museux, J.-M. & others (), 'Handbook on Precision Requirements and Variance Estimation for ESS Household Surveys'.

[4] Azor Martínez, Gerardo; Jiménez Llorente, J. V. P. A. C. & Porras Puga, J. (2011), Study of variance estimation methods in the Spanish Labour Force Survey (EPA)

[5] Behar, R. & Yepes, M. (1991), 'Sobre algunas técnicas de remuestreo: El método de jackknife', Heuristica 5(6), 49–58.

[6] Bellhouse DR (1985) Computing Methods for Variance Estimation in Complex Surveys. Journal of Official Statistics. Vol.1, No.3, 1985

[7] Berger, Y.G. (2004), A Simple Variance Estimator for Unequal Probability Sampling Without Replacement. Journal of Applied Statistics, 31, 305-315.

[8] Berger, Y. G. (2008), 'A note on the asymptotic equivalence of jackknife and linearization variance estimation for the Gini coefficient'.

[9] Binder, David A. (1983). On the variances of asymptotically normal estimators from complex surveys. International Statistical Review, 51, 279-292.

[10] Bruch, C.; Münnich, R. & Zins, S. (2011), 'Variance estimation for complex surveys', Deliverable D3 1.

[11] Canty, A. J. & Davison, A. C. (1999), 'Resampling-based Variance Estimation for Labour Force Surveys', Journal of the Royal Statistical Society: Series D (The Statistician) 48(3), 379–391.

[12] Cochran, W. G. (1977), Sampling Techniques: 3rd Ed, J. Wiley.

[13] Deville, J.-C. & Särndal, C.-E. (1992), 'Calibration estimators in survey sampling', Journal of the American statistical Association 87(418), 376–382.

[14] Deville, J.-C.; Särndal, C.-E. & Sautory, O. (1993), 'Generalized raking procedures in survey sampling', Journal of the American statistical Association 88(423), 1013–1020.

[15] Efron, B. & Tibshirani, R. J. (1994), An introduction to the bootstrap, CRC press.

[16] Fay, R. E. & Train, G. (1995), Aspects of survey and model-based postcensal estimation of income and poverty characteristics for states and counties, in 'Proceedings of the Section on Government Statistics', pp. 154–159.

[17] Girard, C. (2009), The Rao-Wu rescaling bootstrap: from theory to practice, in 'Federal Committee on Statistical Methodology Research Conference', pp. 2–4.

[18] Guillen, Montserrat, M. X. (1996), 'Estimació de la variància mostral a l'Enquesta de Població Activa', Questio Quaderns d Estadistica, Sistemes, Informatica i Investigacio Operativa 20(2), 259–272.

[19] Haziza, D.; Thompson, K. J. & Yung, W. (2010), 'The effect of nonresponse adjustments on variance estimation', Survey Methodology 36(1), 35–43.

[20] INE (2012), 'Diseño de la Encuesta y Evaluación de la calidad de los datos. Informe Técnico.', Technical report, Intituto Nacional de Estadística.

[21] Judkins, D. R. (1990), 'Fay's method for variance estimation', Journal of Official Statistics 6(3), 223–239.
[22] Kerns, G. J. (2010), Introduction to probability and statistics using r.
[23] Kleim, G. & Bélanger, Y. (2007), Using bootstrap variance calculations for a survey with a simple design: The case of the 2005 National Survey of the Work and Health of Nurses, in 'Joint Statistical Meetings, Section on Survey Research Methods'.
[24] Kolenikov, S. & others (2008), Survey bootstrap and bootstrap weights, in 'Summer North American Stata Users' Group Meetings'.
[25] Kott, P. S. (2001), 'Using the delete-a-group jackknife variance estimator in NASS surveys'.
[26] Krewski, D. & Rao, J. (1981), 'Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods', The Annals of Statistics, 1010–1019.
[27] Mach, L.; Saidi, A. & Pettapiece, R. (2007), Study of the properties of the Rao-Wu bootstrap variance estimator: what happens when assumptions do not hold?, in 'Proceedings of the Survey Methods Section, SSC Annual Meeting'.
[28] Martín, X. & Guillen, M. (1992), Estimación de totales y su error de muestreo en la Encuesta de Población Activa. Document de Treball.
[29] Mayor Gallego, J. A. (1997), 'Estimadores de Razón: Una revisión', Qüestiió 21(1), 109-149.
[30] Mirás Amor, J. (1976), 'Estimación de errores de muestreo', Estadística Española(72-73), 7–20.
[31] Mirás Amor, J. (1977), 'Estimación de errores de muestreo II - Método de las semimuestras reiteradas', Estadística Española (74-75), 41–60.
[32] Palmer Arrache, Catalina and Eslava Gómez, Guillermina, I. M. R. (2001), 'Método de remuestreo para el cálculo de varianzas en muestreos complejos: aplicación a la enal'96', Monografías - Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas. UNAM 10(25), 1-120.
[33] Pérez Arriero, C. (2008), 'Calibrating a household survey by using the CALMAR program', Boletín de Estadística e Investigación Operativa 24(2), 22-28.
[34] Popiński, W. (2006), 'Development of the Polish Labour Force Survey', Statistics in Transition 5(7).
[35] Preston, J. (2009), 'Rescaled bootstrap for stratified multistage sampling', Survey Methodology 35(2), 227–234.
[36] Rao, J. (1994), Resampling methods for complex surveys, in 'Proceedings of the 1994 Joint Statistical Meetings, Survey Research Methods Section, American Statistical Association', pp. 35–41.
[37] Rao, J. & Tausi, M. (2004), 'Estimating function jackknife variance estimators under stratified multistage sampling', Communications in Statistics-Theory and Methods 33(9), 2087–2095.
[38] Rao JNK, Yung W, Hidiroglou MA (2002) Estimating equations for the analysis of survey data using poststratification information. Sankhya 64 Series A Part 2, 364-378.
[39] Sarndal C-E, Swensson B, Wretman J (1991) Model Assisted Survey Sampling. Springer.
[40] SAUTORY, O. (1993). La macro CALMAR. Paris: INSEE.
[41] Survey R package: `http://cran.r-project.org/web/packages/survey/survey.pdf`.
[42] Valliant, R. (1993), 'Poststratification and conditional variance estimation', Journal of the American Statistical Association 88(421), 89–96.
[43] Van Kerm, P. (2013), Repeated half-sample bootstrap resampling, in 'United Kingdom Stata Users' Group Meetings 2013'.
[44] Wolter, J.; Volz, R. & Woo, A. (1985), 'Automatic generation of gripping positions', #IEEE_J_SMC# SMC-15(2), 204–213.

Apèndix A
Codi R - Dades

```
require(sas7bdat)
epa113<-read.sas7bdat('epa113.sas7bdat')
epa213<-read.sas7bdat('epa213.sas7bdat')
epa313<-read.sas7bdat('epa313.sas7bdat')
epa413<-read.sas7bdat('epa413.sas7bdat')
epa114<-read.sas7bdat('epa114.sas7bdat')

#save(epa113,epa213,epa313,epa413,epa114,epa214,epa314,
    epa414,file="0_RAWS.RData")
save(epa113,epa213,epa313,epa413,file="0_13.RData")

epa114<-read.sas7bdat('epa114.sas7bdat')
epa214<-read.sas7bdat('epa214.sas7bdat')
epa314<-read.sas7bdat('epa314.sas7bdat')
epa414<-read.sas7bdat('epa414.sas7bdat')
detach("package:sas7bdat", unload=TRUE)
save(epa114,epa214,epa314,epa414,file="0_14.RData")

rm(list=ls())
load("0_13.RData")
#load("0_14.RData")

# Proc?s epa113, cal fer-ho per a tots els trimestres
df<-epa414 #posar tots

#tria de variables
df<-df[,c(8:11,13,14,16,50,52,66,72)]

colnames(df)<-c("PROV","STRAT","CENS","HId","PId","AGE","
    SEX","SIT","NAT","NACE","FACT")
```

```
# Mida de la llar (abans de fer cap subset)
df$PId<-as.numeric(df$PId)
dfsize<-aggregate(df$PId,by=list(df$HId),max)
names(dfsize)<-c("HId","SIZE")
dfsize$SIZE[which(dfsize$SIZE>5)]<-5
dfsize$SIZE<-as.factor(dfsize$SIZE)
df<-merge(df,dfsize,by="HId")
rm(dfsize)
df$PId<-as.factor(df$PId)
df$HId<-as.factor(df$HId)

#Sector econ?mic
df$NACE<-as.numeric(as.vector(df$NACE))
df$ECO<-cut(df$NACE,breaks=c(0,4,40,44,100))
levels(df$ECO)<-c("Agric","Indus","Cons","Serv")
df$NACE<-NULL

#GRUPS d'EDAT, variable creada gEDAT
df=subset(df,df$AGE>15)
glabels=0:10
glabels2=0:2
df$gAGE=cut(df$AGE,c
   (15,19,24,29,34,39,44,49,54,59,64,110),include.lowest
   =T,labels=glabels)
df$gAGE2=cut(df$AGE,c(15,29,49,110),include.lowest=T,
   labels=glabels2)
df$SEXgAGE<-interaction(df$SEX,df$gAGE)
df$SEXgAGE2<-interaction(df$SEX,df$gAGE2)
df$SIT<-droplevels(df$SIT)
rm(glabels);rm(glabels2)

#Creacio variable seccio "ordenada":
census<-data.frame(table(df$CENS,df$STRAT,df$PROV))
census<-census[census$Freq>0,];census$Freq<-NULL
names(census)<-c("CENS","STRAT","PROV")
census$CENSOR<-1:nrow(census)
df<-merge(df,census, by=c("PROV","STRAT","CENS"))
rm(census)

#VARIABLE SIT: 3+4 ocupats, 5 aturats 1a feina, 6 aturats
    que han treballat abans, 7+8+9 inactius
df$ACTIN<-as.factor(round((12-as.numeric(as.vector(df$SIT
   )))/10,0)) #1 ACT, 0 INAC
df$EMPUN<-df$SIT
```

```
levels(df$EMPUN)<-c("1","1","2","2","0","0","0") #1:
   ocupat, 2:aturat (0 inactiu)
df$JLESS<- df$SIT
levels(df$JLESS)<-c("0","0","1","2","0","0","0")#1:
   aturats q mai han treballat 2: que s? que ho han fet
   (0:ocupats i inactius)

#ALTRES VARIABLES AFEGIDES MES TARD
df$UNEMP<-df$EMPUN #UNEMP PREN VALOR 0 SI EMPLEAT, 1 SI
   EMPLEAT i NA SI INACTIU
df$UNEMP[which(df$UNEMP==0)]<-NA
df$EMPOP<-df$EMPUN #employment-to-population
df$EMPOP[which(df$EMPOP!=1)]<-0
df$AGE4<-findInterval(df$AGE,c(16,20,25,55,115))
df$EDU<-epac$ESTUD2
df$EDUC<-recode(df$EDU, "c('11','12','80')='1'")
df$EDUC<-recode(df$EDUC, "c('21','22','23')='2'")
df$EDUC<-recode(df$EDUC, "c('31','32','33','34','41')='3'
   ")
df$EDUC<-recode(df$EDUC, "c
   ('51','52','53','54','55','56','61')='4'")

df$BRAN<-cut(as.numeric(as.character(epac$ACT)),breaks=c
   (0,3,9,18,22,25,33,39,43,46,47,53,56,63,66,75,82,84,85,88,99)
   )
require(car)
df$BRAN<-recode(df$BRAN, "c('(3,9]','(33,39]')='(3,9]'")
levels(df$BRAN)=c("1","4","5","6","2",as.character(7:19),
   "3")

epa4140<-df  # FINS AQUI EL TRACTAMENT DE CADA FITXER (
   posar tots)

)
save(epa1130,epa2130,epa3130,epa4130,file="1_13.RData")
save(epa1140,epa2140,epa3140,epa4140,file="1_14.RData")

rm(df)
```

Apèndix B
Codi R - Descriptius

```
#Diagrama de sectors - educacio
par(mar=c(3,5,5,3))
pie(table(epa1$EDUC),col=col4,labels=c("Analfabets i
    educacio primaria","Educacio secundaria 1a. etapa","
    Educacio secundaria 2a. etapa","Educacio superior"),
    cex=0.8, main="Nivell d'educacio (mostra)")
vals=paste(round(prop.table(table(epa1$EDUC))*100,2),"%")
text(0.3,0.4,vals[1])
text(-0.4,0.2,vals[2])
text(-0.3,-0.4,vals[2])
text(0.35,-0.4,vals[4])

#Diagrama de sectors - sectors economics
par(mar=c(3,4,5,5))
vals2=paste(round(prop.table(table(epa1$ECO))*100,2),"%")
pie(table(epa1$ECO),col=col4,labels=paste(vals2," - ",c("
    Agricultura","Industria","Construccio","Serveis")),
    main="Sectors d'activitat (mostra)")

#DISTRIBUCIO POBLACIO SEXE EDAT:
colAd<-adjustcolor("black",alpha=0.5)
plot(df$SEXgAGE,col=c("tomato","mediumspringgreen"),xlab=
    "Age Group",ylab="No. People",main=paste("
    Distribution by sex and age (T",trim,"-20",any,")",
    sep=""),names=rep(0:10,each=2))
legend(1,2050,c("Men   (sample)","Women (sample)","
    Population"),lty=c(1,1,1),lwd=c(4,4,4),col=c("tomato"
    ,"mediumspringgreen",colAd))# Nombre de seccions per
    prov?ncia/estrat (MOSTRA)
#poblacional:
FE=sum(as.vector(aggregate(FACT~SEXgAGE,data=df,FUN=sum))
    [,2])/nrow(df)
```

```
SEXgAGE13<-as.vector(aggregate(FACT~SEXgAGE,data=df,FUN=
    sum))
SEXgAGE13[,2]<-SEXgAGE13[,2]/FE
#plot(SEXgAGE13[,2],col=c("tomato","mediumspringgreen"),
    xlab="Age Group",ylab="# People",main=paste("
    Distribution of population by sex and age (T",trim
    ,"-20",any,")",sep=""),names=rep(0:10,each=2))
#legend(1,1950,c("Men","Women"),lty=c(1,1),lwd=c(4,4),col
    =c("tomato","mediumspringgreen"))# Nombre de seccions
     per prov?ncia/estrat (POBLACI?)
SEXgAGEt13=cbind(SEXgAGE13,as.vector(table(df$SEXgAGE)))
colnames(SEXgAGEt13)[3]="FACTs"
barplot(SEXgAGE13[,2],add=T,col=colAd) #Especificada a
    sota

#Seccions censals a la mostra
Kh<-aggregate(CENS~STRAT+PROV,data=unique(df[,vars]),
    length)
M<-as.data.frame(matrix(c(rep(1:9,4),rep("08",9),rep("17"
    ,9),rep("25",9),rep("43",9),rep(0,36)),nrow=36,ncol
    =3))
colnames(M)<-c("STRAT","PROV","CENS")
M$CENS=rep(0,36)
for(j in 1:26){
  for (i in 1:36){
    if(M[i,1]==Kh[j,1]){
      if(M[i,2]==Kh[j,2]){
        M[i,3]=as.numeric(Kh[j,3])
      }
    }
  }
}
par(mfrow=c(2,2))
par(mar=c(2,2,2,2))
ylim<-c(0,max(M$CENS))
a<-split(M,M$PROV)
sapply(a,FUN=function(x)barplot(x$CENS,names.arg=x$STRAT,
    ylim=ylim,col="firebrick2"))
title("Nombre␣de␣seccions␣censals␣per␣estrat␣a␣la␣mostra"
    , line = -13.75, outer = TRUE)
mtext("Barcelona␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣
    ␣␣Girona", side = 3, line = -3.5, outer = TRUE)
mtext("␣␣␣␣␣␣␣Lleida␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣
    ␣␣␣␣␣␣␣␣␣Tarragona", side = 3, line = -16.5, outer =
    TRUE)
```

```
par(mfrow=c(1,1))
rm(a)

#LLARS PER ESTRAT
par(mfrow=c(2,2))
ylim<-c(0,max(M$HId))
a<-split(M,M$PROV)
sapply(a,FUN=function(x)barplot(x$HId,names.arg=x$STRAT,
   ylim=ylim,col="deepskyblue"))
title("Nombre␣de␣llars␣per␣estrat␣a␣la␣mostra", line =
   -13.5, outer = TRUE)
mtext("Barcelona␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣
   ␣␣Girona", side = 3, line = -3.5, outer = TRUE)
mtext("␣␣␣␣␣␣␣Lleida␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣
   ␣␣␣␣␣␣␣␣␣Tarragona", side = 3, line = -16.5, outer =
   TRUE)
par(mfrow=c(1,1))
rm(a)

#LLARS PER SECCIO
vars<-c("PROV","STRAT","CENS","HId")
vjh<-aggregate(HId~CENS+STRAT+PROV,length,data=unique(df
   [,vars]))
#a:
plot(HId~PROV,data=vjh,col="lawngreen",main=paste("Llars␣
   per␣seccio␣a␣la␣mostra␣efectiva"),names=c("Barcelona"
   ,"Girona","Lleida","Tarragona"))
#posar-ho a l'etiqueta, que es per prov.
summary(vjh$HId)
#b:
par(mfrow=c(2,2))
ylim<-c(0,max(vjh$HId))
a<-split(vjh,vjh$PROV)
sapply(a,FUN=function(x)boxplot(HId~STRAT,data=x,ylim=
   ylim,col="lawngreen"))
title("Llars␣per␣seccio␣censal␣i␣estrat", line = -15.95,
   outer = TRUE)
mtext("Barcelona␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣
   ␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣Girona", side = 3, line = -11,
    outer = TRUE)
mtext("␣␣␣␣␣␣Lleida␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣
   ␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣Tarragona", side = 3, line
    = -19.5, outer = TRUE)
par(mfrow=c(1,1))
rm(a);rm(vars)
```

```
#POBLACIO PER ESTRAT
par(mfrow=c(2,2))
ylim<-c(0,max(M$PId)+)
a<-split(M,M$PROV)
sapply(a,FUN=function(x)boxplot(HId~STRAT,data=x,ylim=
   ylim,col="lawngreen"))
sapply(a,FUN=function(x)barplot(PId,~STRAT,ylim=ylim,col=
   "darkgoldenrod1"))
title("Nombre␣de␣persones␣per␣estrat", line = -13.5,
   outer = TRUE)
mtext("Barcelona␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣
   ␣␣Girona", side = 3, line = -3.5, outer = TRUE)
mtext("␣␣␣␣␣␣␣Lleida␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣
   ␣␣␣␣␣␣␣␣␣␣Tarragona", side = 3, line = -16.5, outer =
   TRUE)
par(mfrow=c(1,1))
rm(a)
```

Apèndix C
Codi R - Comparació dels factors

```
factors<-data.frame(df$PROV,df$STRAT,df$CENS,df$HId,INE=
   df$FACT,HT=df$facD,rao=df$facP,calib=weights(cal))
head(factors);factors<-unique(factors)
head(factors);factors<-factors[,-(1:4)]
col4=c("lawngreen","darkgoldenrod1","firebrick2","
   deepskyblue")
boxplot(factors,col=col4,main="Comparacio dels factors
   (1-2013)",names=c("INE","H - T","Rao separat", "
   Calibrat"))
summary(factors)
cor(factors)
xtable(cor(factors),caption=rep("Correlaci? dels pesos de
   disseny estimats",2),label="Tcorrelacio",digits=3,
   align="|l|cccc|")
ylim=c(0,max(factors))
plot(factors$INE,type="l",ylim=ylim,main=paste("Factors d
   'elevacio",sep=""),xlab="Llar",ylab="Factor d'
   elevacio")
colAd1<-adjustcolor("deepskyblue",alpha.f=0.5)
points(factors$calib,col=colAd1,type="l",lwd=1,ylim=ylim)
points(factors$HT,col="darkgoldenrod1",type="l",lwd=2,
   ylim=ylim)
points(factors$rao,col="firebrick2",type="l",lwd=2,ylim=
   ylim)
legend("topright",c("INE","H-T","Rao s.","Calibrat"),lty=
   c(1,1,1,1),lwd=c(4,4,4,4),col=c(1,"darkgoldenrod1","
   firebrick2","deepskyblue"))
plot(factors,pch=16,cex=0.3)
par(mfrow=c(1,3))
qqplot(factors$INE,factors$calib,pch=16,cex=0.5)
qqplot(factors$INE,factors$HT,pch=16,cex=0.5)
qqplot(factors$INE,factors$rao,pch=16,cex=0.5)
```

```
par(mfrow=c(2,2))
par(mar=c(5,2,5,2))
hist(factors$INE,ylim=ylim);hist(factors$HT,ylim=ylim);
   hist(factors$rao,ylim=ylim);hist(factors$calib,ylim=
   ylim)
par(mfrow=c(1,1))

#test de normalitat Shapiro-Wilk. p-value < 2.2e-16
m <- mean(factors$INE)
s <- sd(factors$INE)
ks.test(factors$INE, "pnorm", m, s)

m2 <- mean(factors$calib)
s2 <- sd(factors$calib)
ks.test(factors$calib, "pnorm", m2, s2)

ks.test(factors$INE,factors$calib)
plot(ecdf(factors$INE),col="firebrick2",lwd=1, main= "
   Funcions␣de␣distribucio")
lines(ecdf(factors$calib),col="deepskyblue",lwd=1)
legend("right",c("Factors␣INE","Factors␣calibrats"),lty=c
   (1,1),lwd=c(4,4),col=c("firebrick2","deepskyblue"))

wilcox.test(factors$INE,factors$calib,paired=T)#rebutjo
wilcox.test(factors$INE,factors$N2,paired=T)
kruskal.test(factors[,c("INE","calib")],paired=T)
rm(factors);rm(ylim)
```

Apèndix D
Codi R - Calibratge

```
EPA0=list(epa1130,epa2130,epa3130,epa4130,epa1140,epa2140
   ,epa3140,epa4140)
EPA=list(epa1131,epa2131,epa3131,epa4131,epa1141,epa2141,
   epa3141,epa4141)
STHT<-list(NULL)
STrao<-list(NULL)
STTay<-list(NULL)
STmrb<-list(NULL)
STJkn<-list(NULL)
STbrr<-list(NULL)
STbrr50<-list(NULL)
STbrr40h<-list(NULL)

for(i in 1:length(EPA)){
  df=EPA[[i]]
  cat("TRIMESTRE",i,"\n")
  ####Estimador de H-T####
  #Quants habitatges hi ha per estrat a la poblacio?
     TENIR EN COMPTE POSTERIORMENT EL #HAB per LLAR
  #Estem utilitzant els pesos de l'INE per tornar a les
     dades de partida.
  Vh = aggregate(FACT~STRAT+PROV,data=unique(df[c("PROV",
     "STRAT","FACT","HId")]),FUN=sum)
  #Quants habitatges hi ha per estrat a la mostra?
  vh = aggregate(HId~STRAT+PROV,data=unique(df[c("PROV","
     STRAT","HId")]),FUN=length)
  #### Factors del disseny:
  vars<-c("PROV","STRAT","CENS")
  Kh<-aggregate(CENS~STRAT+PROV,data=unique(df[,vars]),
     length)
  facD = cbind(Vh[,1:2],Vh[,3]/(18*Kh$CENS));colnames(
     facD)[3]<-"facD";facD
```

```
df<-merge(df,facD,by=c("PROV","STRAT"))
svdd<-svydesign(ids=~CENS+HId,strata=~droplevels(
    interaction(PROV,STRAT)),data=df,weights=~facD,nest
    =T)
STHT[[i]]<-svytotal(~ACTIN+EMPUN+JLESS,svdd,deff=T)
STHT[[i]]<-cbind(data.frame(STHT[[i]]),confint(STHT[[i
    ]]),cv(STHT[[i]])*100)[c(2,3,4,7,8,1),c
    (1,2,4,5,6,3)]
colnames(STHT[[i]])[5]<-"CV(%)";rownames(STHT[[i]])<-c(
    "Actius","Ocupats","Aturats","At.␣no␣treb.","At.␣si
    ␣treb.","Inactius")
print("Estimador␣H-T")
print(STHT[[i]])

####Estimador de rao separat####
facP<-cbind(aggregate(FACT~STRAT+PROV,data=df,FUN=sum)
    [1:2],aggregate(FACT~STRAT+PROV,data=df,FUN=sum)$
    FACT/aggregate(PId~STRAT+PROV,data=df,FUN=length)$
    PId);colnames(facP)[3]<-"facP";facP
df<-merge(df,facP,by=c("PROV","STRAT"))
#Semblanca amb els anteriors
cor(facP[,3],facD[,3])
svd<-svydesign(ids=~CENS+HId,strata=~droplevels(
    interaction(PROV,STRAT)),data=df,weights=~facP,nest
    =T)
STrao[[i]]<-svytotal(~ACTIN+EMPUN+JLESS,svd,deff=T)
STrao[[i]]<-cbind(data.frame(STrao[[i]]),confint(STrao
    [[i]]),cv(STrao[[i]])*100)[c(2,3,4,7,8,1),c
    (1,2,4,5,6,3)]
colnames(STrao[[i]])[5]<-"CV(%)";rownames(STrao[[i]])<-
    c("Actius","Ocupats","Aturats","At.␣no␣treb.","At.␣
    si␣treb.","Inactius")
print("Estimador␣de␣rao␣separat")
print(STrao[[i]])

####Estimador calibrat a partir dels de rao####
#Partim de l'estimador de rao separat perque corregeix
    d'alguna manera el biaix: elevacions facP
#El disseny amb els factors de poblacio (era svd).
#El de l'INE per calcular els poblacionals
svdine<-svydesign(ids=~CENS+HId,strata=~droplevels(
    interaction(PROV,STRAT)),data=df,weights=~FACT,nest
    =T)
#"Fonts externes":
pop.sex.age<-as.vector(svytable(~SEXgAGE,svdine))
```

```
pop.prov <-as.vector(svytable(~PROV,svdine))
pop.nat<-as.vector(svytable(~NAT,svdine))
#REMOSTREIG
#MRBweights(Bootstrap)
mrb<-as.svrepdesign(svd,type="mrbbootstrap",replicates
   =50,mse=T)
#JKnweights(Jacknife)
jkn<-as.svrepdesign(svd,type="JKn",mse=T,compress=F)
#BRR (semimostres reiterades balancejades)
fay<-as.svrepdesign(svd,type="Fay",mse=T,fay.rho=0.75,
   compress=F,small="merge",large="merge")
brr<-as.svrepdesign(svd,type="BRR",mse=T,compress=F,
   small="merge",large="merge")
brr50<-as.svrepdesign(svd,type="BRR",mse=T,compress=F,
   fay.rho=0.5,small="merge",large="merge")
ncol(brr$repweights) # NOMBRE DE RePLIQUES: MiNIM
   MULTIPLE DE 4 SUPERIOR AL NOMBRE D'ESTRATS
#BRR: Per obtenir mes repliques cal passar una matriu
   de Hadamard d'ordre superior
brr40h<-as.svrepdesign(svd,type="BRR",hadamard.matrix=
   hadamard(39),mse=T,compress=T,small="merge",large="
   merge")

if(i>4){
  pop.size<-as.vector(svytable(~SIZE,svdine))
  pop.sex.age2<-as.vector(svytable(~SEXgAGE2,svdine))
  #Calibratge:
  calfay<-calibrate(fay,~SEXgAGE+NAT+PROV+SIZE+SEXgAGE2
     -1,c(pop.sex.age,pop.nat[-1],pop.prov[-1],pop.
     size[-1],pop.sex.age2[-1]),calfun="linear",bounds
     =c(0.1,10),aggregate.index=~HId,compress=F)
  cal<-calibrate(svd,~SEXgAGE+NAT+PROV+SIZE+SEXgAGE2-1,
     c(pop.sex.age,pop.nat[-1],pop.prov[-1],pop.size
     [-1],pop.sex.age2[-1]),calfun="linear",bounds=c
     (0.1,10),aggregate=2)
  calmrb<-calibrate(mrb,~SEXgAGE+NAT+PROV+SIZE+SEXgAGE2
     -1,c(pop.sex.age,pop.nat[-1],pop.prov[-1],pop.
     size[-1],pop.sex.age2[-1]),calfun="linear",bounds
     =c(0.1,10),aggregate.index=~HId,compress=F)
  caljkn<-calibrate(jkn,~SEXgAGE+NAT+PROV+SIZE+SEXgAGE2
     -1,c(pop.sex.age,pop.nat[-1],pop.prov[-1],pop.
     size[-1],pop.sex.age2[-1]),calfun="linear",bounds
     =c(0.1,10),aggregate.index=~HId,compress=F)
```

```
  calbrr<-calibrate(brr,~SEXgAGE+NAT+PROV+SIZE+SEXgAGE2
     -1,c(pop.sex.age,pop.nat[-1],pop.prov[-1],pop.
     size[-1],pop.sex.age2[-1]),calfun="linear",bounds
     =c(0.1,10),aggregate.index=~HId,compress=F)
  calbrr50<-calibrate(brr50,~SEXgAGE+NAT+PROV+SIZE+
     SEXgAGE2-1,c(pop.sex.age,pop.nat[-1],pop.prov
     [-1],pop.size[-1],pop.sex.age2[-1]),calfun="
     linear",bounds=c(0.1,10),aggregate.index=~HId,
     compress=F)
  calbrr40h<-calibrate(brr40h,~SEXgAGE+NAT+PROV+SIZE+
     SEXgAGE2-1,c(pop.sex.age,pop.nat[-1],pop.prov
     [-1],pop.size[-1],pop.sex.age2[-1]),calfun="
     linear",bounds=c(0.1,10),aggregate.index=~HId,
     compress=F)
}else{
  #Calibratge:
  calfay<-calibrate(fay,~SEXgAGE+NAT+PROV-1,c(pop.sex.
     age,pop.nat[-1],pop.prov[-1]),calfun="linear",
     bounds=c(0.1,10),aggregate=2)
  cal<-calibrate(svd,~SEXgAGE+NAT+PROV-1,c(pop.sex.age,
     pop.nat[-1],pop.prov[-1]),calfun="linear",bounds=
     c(0.1,10),aggregate=2)
  calmrb<-calibrate(mrb,~SEXgAGE+NAT+PROV-1,c(pop.sex.
     age,pop.nat[-1],pop.prov[-1]),calfun="linear",
     bounds=c(0.1,10),aggregate.index=~HId,compress=F)
  caljkn<-calibrate(jkn,~SEXgAGE+NAT+PROV-1,c(pop.sex.
     age,pop.nat[-1],pop.prov[-1]),calfun="linear",
     bounds=c(0.1,10),aggregate.index=~HId,compress=F)
  calbrr<-calibrate(brr,~SEXgAGE+NAT+PROV-1,c(pop.sex.
     age,pop.nat[-1],pop.prov[-1]),calfun="linear",
     bounds=c(0.1,10),aggregate.index=~HId,compress=F)
  calbrr50<-calibrate(brr50,~SEXgAGE+NAT+PROV-1,c(pop.
     sex.age,pop.nat[-1],pop.prov[-1]),calfun="linear"
     ,bounds=c(0.1,10),aggregate.index=~HId,compress=F
     )
  calbrr40h<-calibrate(brr40h,~SEXgAGE+NAT+PROV-1,c(pop
     .sex.age,pop.nat[-1],pop.prov[-1]),calfun="linear
     ",bounds=c(0.1,10),aggregate.index=~HId,compress=
     F)
}
#BRR Fay 0.5
STfay[[i]]<-svytotal(~ACTIN+EMPUN+JLESS,calfay,deff=T)
STfay[[i]]<-cbind(data.frame(STfay[[i]]),confint(STfay
    [[i]]),cv(STfay[[i]])*100)[c(2,3,4,7,8,1),c
    (1,2,4,5,6,3)]
```

```
colnames(STfay[[i]])[5]<-"CV(%)";rownames(STfay[[i]])<-
    c("Actius","Ocupats","Aturats","At.␣no␣treb.","At.␣
    si␣treb.","Inactius")
print("Estimador␣calibrat␣BRR␣Fay␣0.5")
print(STfay[[i]])
#Taylor
STTay[[i]]<-svytotal(~ACTIN+EMPUN+JLESS,cal,deff=T)
STTay[[i]]<-cbind(data.frame(STTay[[i]]),confint(STTay
    [[i]]),cv(STTay[[i]])*100)[c(2,3,4,7,8,1),c
    (1,2,4,5,6,3)]
colnames(STTay[[i]])[5]<-"CV(%)";rownames(STTay[[i]])<-
    c("Actius","Ocupats","Aturats","At.␣no␣treb.","At.␣
    si␣treb.","Inactius")
print("Estimador␣calibrat␣Taylor")
print(STTay[[i]])
#Bootstrap
STmrb[[i]]<-svytotal(~ACTIN+EMPUN+JLESS,calmrb,deff=T)
STmrb[[i]]<-cbind(data.frame(STmrb[[i]]),confint(STmrb
    [[i]]),cv(STmrb[[i]])*100)[c(2,3,4,7,8,1),c
    (1,2,4,5,6,3)]
colnames(STmrb[[i]])[5]<-"CV(%)";rownames(STmrb[[i]])<-
    c("Actius","Ocupats","Aturats","At.␣no␣treb.","At.␣
    si␣treb.","Inactius")
print("Estimador␣calibrat␣Bootstrap")
print(STmrb[[i]])
#Jackknife
STJkn[[i]]<-svytotal(~ACTIN+EMPUN+JLESS,caljkn,deff=T)
STJkn[[i]]<-cbind(data.frame(STJkn[[i]]),confint(STJkn
    [[i]]),cv(STJkn[[i]])*100)[c(2,3,4,7,8,1),c
    (1,2,4,5,6,3)]
colnames(STJkn[[i]])[5]<-"CV(%)";rownames(STJkn[[i]])<-
    c("Actius","Ocupats","Aturats","At.␣no␣treb.","At.␣
    si␣treb.","Inactius")
print("Estimador␣calibrat␣Jackknife")
print(STJkn[[i]])
#BRR
STbrr[[i]]<-svytotal(~ACTIN+EMPUN+JLESS,calbrr,deff=T)
STbrr[[i]]<-cbind(data.frame(STbrr[[i]]),confint(STbrr
    [[i]]),cv(STbrr[[i]])*100)[c(2,3,4,7,8,1),c
    (1,2,4,5,6,3)]
colnames(STbrr[[i]])[5]<-"CV(%)";rownames(STbrr[[i]])<-
    c("Actius","Ocupats","Aturats","At.␣no␣treb.","At.␣
    si␣treb.","Inactius")
print("Estimador␣calibrat␣BRR")
print(STbrr[[i]])
```

```
  #BRR Fay 0.5
  STbrr50[[i]]<-svytotal(~ACTIN+EMPUN+JLESS,calbrr50,deff
     =T)
  STbrr50[[i]]<-cbind(data.frame(STbrr50[[i]]),confint(
     STbrr50[[i]]),cv(STbrr50[[i]])*100)[c(2,3,4,7,8,1),
     c(1,2,4,5,6,3)]
  colnames(STbrr50[[i]])[5]<-"CV(%)";rownames(STbrr50[[i
     ]])<-c("Actius","Ocupats","Aturats","At. no treb.",
     "At. si treb.","Inactius")
  print("Estimador calibrat BRR Fay 0.5")
  print(STbrr50[[i]])
  #BRR40
  STbrr40h[[i]]<-svytotal(~ACTIN+EMPUN+JLESS,calbrr40h,
     deff=T)
  STbrr40h[[i]]<-cbind(data.frame(STbrr40h[[i]]),confint(
     STbrr40h[[i]]),cv(STbrr40h[[i]])*100)[c
     (2,3,4,7,8,1),c(1,2,4,5,6,3)]
  colnames(STbrr40h[[i]])[5]<-"CV(%)";rownames(STbrr40h[[
     i]])<-c("Actius","Ocupats","Aturats","At. no treb."
     ,"At. si treb.","Inactius")
  print("Estimador calibrat BRR 40")
  print(STbrr40h[[i]])
  cat("\n")
}
```

Apèndix E
Codi R - Comparació dels mètodes

1. Variabilitat entre estimacions

Només posem el codi per un trimestre de 2014, amb un calibratge més complex.

```
P7TTay=list(NULL)
P7Tbrr=list(NULL)
P7TJkn=list(NULL)
P7Tmrb=list(NULL)
df=epa3140
  ####Estimador de rao separat####
  facP<-cbind(aggregate(FACT~STRAT+PROV,data=df,FUN=sum)
     [1:2],aggregate(FACT~STRAT+PROV,data=df,FUN=sum)$
     FACT/aggregate(PId~STRAT+PROV,data=df,FUN=length)$
     PId);colnames(facP)[3]<-"facP";facP
  df<-merge(df,facP,by=c("PROV","STRAT"))
  svd<-svydesign(ids=~CENS+HId,strata=~droplevels(
     interaction(PROV,STRAT)),data=df,weights=~facP,nest
     =T)
  ####Estimador calibrat a partir dels de rao####
  svdine<-svydesign(ids=~CENS+HId,strata=~droplevels(
     interaction(PROV,STRAT)),data=df,weights=~FACT,nest
     =T)
  #"Fonts externes":
  pop.sex.age<-as.vector(svytable(~SEXgAGE,svdine))
  pop.prov <-as.vector(svytable(~PROV,svdine))
  pop.nat<-as.vector(svytable(~NAT,svdine))
  pop.size<-as.vector(svytable(~SIZE,svdine))
  pop.sex.age2<-as.vector(svytable(~SEXgAGE2,svdine))

for(i in 1:50){
  cat("PROVA",i,"\n")
  #REMOSTREIG
```

```
#MRBweights(Bootstrap)
mrb<-as.svrepdesign(svd,type="mrbbootstrap",replicates
   =100,mse=T)
#JKnweights(Jacknife)
jkn<-as.svrepdesign(svd,type="JKn",mse=T,compress=F)
#BRR (semimostres reiterades balancejades)
brr<-as.svrepdesign(svd,type="BRR",mse=T,compress=F,
   small="merge",large="merge")
cal<-calibrate(svd,~SEXgAGE+NAT+PROV+SIZE+SEXgAGE2-1,c(
   pop.sex.age,pop.nat[-1],pop.prov[-1],pop.size[-1],
   pop.sex.age2[-1]),calfun="linear",bounds=c(0.1,10),
   aggregate=2)
calmrb<-calibrate(mrb,~SEXgAGE+NAT+PROV+SIZE+SEXgAGE2
   -1,c(pop.sex.age,pop.nat[-1],pop.prov[-1],pop.size
   [-1],pop.sex.age2[-1]),calfun="linear",bounds=c
   (0.1,10),aggregate.index=~HId,compress=F)
caljkn<-calibrate(jkn,~SEXgAGE+NAT+PROV+SIZE+SEXgAGE2
   -1,c(pop.sex.age,pop.nat[-1],pop.prov[-1],pop.size
   [-1],pop.sex.age2[-1]),calfun="linear",bounds=c
   (0.1,10),aggregate.index=~HId,compress=F)
calbrr<-calibrate(brr,~SEXgAGE+NAT+PROV+SIZE+SEXgAGE2
   -1,c(pop.sex.age,pop.nat[-1],pop.prov[-1],pop.size
   [-1],pop.sex.age2[-1]),calfun="linear",bounds=c
   (0.1,10),aggregate.index=~HId,compress=F)
#Taylor
P7TTay[[i]]<-svytotal(~ACTIN+EMPUN+JLESS,cal,deff=T)
P7TTay[[i]]<-cbind(data.frame(P7TTay[[i]]),confint(
   P7TTay[[i]]),cv(P7TTay[[i]])*100)[c(2,3,4,7,8,1),c
   (1,2,4,5,6,3)]
colnames(P7TTay[[i]])[5]<-"CV(%)";rownames(P7TTay[[i]])
   <-c("Actius","Ocupats","Aturats","At.␣no␣treb.","At
   .␣si␣treb.","Inactius")
print("Estimador␣calibrat␣Taylor")
print(P7TTay[[i]])
#Bootstrap
P7Tmrb[[i]]<-svytotal(~ACTIN+EMPUN+JLESS,calmrb,deff=T)
P7Tmrb[[i]]<-cbind(data.frame(P7Tmrb[[i]]),confint(
   P7Tmrb[[i]]),cv(P7Tmrb[[i]])*100)[c(2,3,4,7,8,1),c
   (1,2,4,5,6,3)]
colnames(P7Tmrb[[i]])[5]<-"CV(%)";rownames(P7Tmrb[[i]])
   <-c("Actius","Ocupats","Aturats","At.␣no␣treb.","At
   .␣si␣treb.","Inactius")
print("Estimador␣calibrat␣Bootstrap")
print(P7Tmrb[[i]])
#Jackknife
```

```
  P7TJkn[[i]]<-svytotal(~ACTIN+EMPUN+JLESS,caljkn,deff=T)
  P7TJkn[[i]]<-cbind(data.frame(P7TJkn[[i]]),confint(
     P7TJkn[[i]]),cv(P7TJkn[[i]])*100)[c(2,3,4,7,8,1),c
     (1,2,4,5,6,3)]
  colnames(P7TJkn[[i]])[5]<-"CV(%)";rownames(P7TJkn[[i]])
     <-c("Actius","Ocupats","Aturats","At.␣no␣treb.","At
     .␣si␣treb.","Inactius")
  print("Estimador␣calibrat␣Jackknife")
  print(P7TJkn[[i]])
  #BRR
  P7Tbrr[[i]]<-svytotal(~ACTIN+EMPUN+JLESS,calbrr,deff=T)
  P7Tbrr[[i]]<-cbind(data.frame(P7Tbrr[[i]]),confint(
     P7Tbrr[[i]]),cv(P7Tbrr[[i]])*100)[c(2,3,4,7,8,1),c
     (1,2,4,5,6,3)]
  colnames(P7Tbrr[[i]])[5]<-"CV(%)";rownames(P7Tbrr[[i]])
     <-c("Actius","Ocupats","Aturats","At.␣no␣treb.","At
     .␣si␣treb.","Inactius")
  print("Estimador␣calibrat␣BRR")
  print(P7Tbrr[[i]])
}

T7R<-matrix(nrow=6,ncol=8)
for(i in 1:6){
  T7R[i,1]=mean(unlist(lapply(P7TTay, `[[`, i,5)));T7R[i
     ,2]=sd(unlist(lapply(P7TTay, `[[`, i,5)))
  T7R[i,3]=mean(unlist(lapply(P7Tmrb, `[[`, i,5)));T7R[i
     ,4]=sd(unlist(lapply(P7Tmrb, `[[`, i,5)))
  T7R[i,5]=mean(unlist(lapply(P7TJkn, `[[`, i,5)));T7R[i
     ,6]=sd(unlist(lapply(P7TJkn, `[[`, i,5)))
  T7R[i,7]=mean(unlist(lapply(P7Tbrr, `[[`, i,5)));T7R[i
     ,8]=sd(unlist(lapply(P7Tbrr, `[[`, i,5)))
}
rownames(T7R)<-c("Actius","Ocupats","Aturats","At.␣no␣
   treb.","At.␣si␣treb.","Inactius")
colnames(T7R)<-c("Taylor","SD","Bootstrap","SD","
   Jackknife","SD","BRR","SD")
```

2. Intervals de confiança

```
require(plotrix)

res3<-data.frame(t(c("METODE","MEAN","SE","CV","L95","U95
   ")),stringsAsFactors=F)
colnames(res3)<-res3[1,];res3<-res3[-1,];res3
```

```
a<-svytotal(~EMPUN, subset(svdine,ACTIN=="1"))
res3[nrow(res3)+1,]<-cbind("INE",a[2],SE(a)[2],cv(a)[2],
   confint(a)[2,1],confint(a)[2,2])

a<-svytotal(~EMPUN, subset(cal,ACTIN=="1"))
res3[nrow(res3)+1,]<-cbind("Taylor",a[2],SE(a)[2],cv(a)
   [2],confint(a)[2,1],confint(a)[2,2])

a<-svytotal(~EMPUN, subset(calbrr,ACTIN=="1"))
res3[nrow(res3)+1,]<-cbind("BRR",a[2],SE(a)[2],cv(a)[2],
   confint(a)[2,1],confint(a)[2,2])

a<-svytotal(~EMPUN, subset(caljkn,ACTIN=="1"))
res3[nrow(res3)+1,]<-cbind("JKn",a[2],SE(a)[2],cv(a)[2],
   confint(a)[2,1],confint(a)[2,2])

a<-svytotal(~EMPUN, subset(calmrb,ACTIN=="1"))
res3[nrow(res3)+1,]<-cbind("MRB",a[2],SE(a)[2],cv(a)[2],
   confint(a)[2,1],confint(a)[2,2])

sapply(2:ncol(res3),FUN=function(x) res3[,x]<<-as.numeric
   (res3[,x]))
ylim<-c(min(res3$L95),max(res3$U95))
plotCI(x = 1:nrow(res3), y = res3$MEAN, li = res3$L95,
   xaxt="n",ui = res3$U95,ylim=ylim,main="Intevals␣de␣
   confianCa␣per␣a␣la␣poblacio␣desocupada",ylab="")
axis(1, at=1:5, labels=res3$METODE)
```

3. Temps

```
tmrb50<-NULL
tmrb100<-NULL
tmrb200<-NULL
tjkn<-NULL
tbrr<-NULL
tbrr50<-NULL
ttay<-NULL
time<-NULL

for (i in 1:50){
  #svd<-svydesign(ids=~CENS+HId,strata=~droplevels(
     interaction(PROV,STRAT)),data=df,weights=~facP,nest
     =T)
ptm=proc.time()
```

```
cal<-calibrate(svd,~SEXgAGE+NAT+PROV+SIZE+SEXgAGE2-1,c(
   pop.sex.age,pop.nat[-1],pop.prov[-1],pop.size[-1],pop
   .sex.age2[-1]),calfun="linear",bounds=c(0.1,10),
   aggregate=2)
ttay[i]=proc.time()-ptm
ptm=proc.time()
  jkn<-as.svrepdesign(svd,type="JKn",mse=T,compress=F)
  caljkn<-calibrate(jkn,~SEXgAGE+NAT+PROV+SIZE+SEXgAGE2
     -1,c(pop.sex.age,pop.nat[-1],pop.prov[-1],pop.size
     [-1],pop.sex.age2[-1]),calfun="linear",bounds=c
     (0.1,10),aggregate.index=~HId,compress=F)
tjkn[i]=proc.time()-ptm
ptm=proc.time()
  mrb<-as.svrepdesign(svd,type="mrbbootstrap",replicates
     =50,mse=T)
  calmrb<-calibrate(mrb,~SEXgAGE+NAT+PROV+SIZE+SEXgAGE2
     -1,c(pop.sex.age,pop.nat[-1],pop.prov[-1],pop.size
     [-1],pop.sex.age2[-1]),calfun="linear",bounds=c
     (0.1,10),aggregate.index=~HId,compress=F)
tmrb50[i]=proc.time()-ptm;time
ptm=proc.time()
  brr<-as.svrepdesign(svd,type="BRR",mse=T,compress=F,
     small="merge",large="merge")
  calbrr<-calibrate(brr,~SEXgAGE+NAT+PROV+SIZE+SEXgAGE2
     -1,c(pop.sex.age,pop.nat[-1],pop.prov[-1],pop.size
     [-1],pop.sex.age2[-1]),calfun="linear",bounds=c
     (0.1,10),aggregate.index=~HId,compress=F)
tbrr[i]=proc.time()-ptm
ptm=proc.time()
  brr50<-as.svrepdesign(svd,type="BRR",mse=T,compress=F,
     fay.rho=0.5,small="merge",large="merge")
  calbrr50<-calibrate(brr50,~SEXgAGE+NAT+PROV+SIZE+
     SEXgAGE2-1,c(pop.sex.age,pop.nat[-1],pop.prov[-1],
     pop.size[-1],pop.sex.age2[-1]),calfun="linear",
     bounds=c(0.1,10),aggregate.index=~HId,compress=F)
tbrr50[i]=proc.time()-ptm
#ptm=proc.time()
  #brr40h<-as.svrepdesign(svd,type="BRR",hadamard.matrix=
     hadamard(39),mse=T,compress=T,small="merge",large="
     merge")
  #calbrr40h<-calibrate(brr40h,~SEXgAGE+NAT+PROV+SIZE+
     SEXgAGE2-1,c(pop.sex.age,pop.nat[-1],pop.prov[-1],
     pop.size[-1],pop.sex.age2[-1]),calfun="linear",
     bounds=c(0.1,10),aggregate.index=~HId,compress=F)
#time[6]=proc.time()-ptm
```

```
ptm=proc.time()
mrb<-as.svrepdesign(svd,type="mrbbootstrap",replicates
   =100,mse=T)
calmrb<-calibrate(mrb,~SEXgAGE+NAT+PROV+SIZE+SEXgAGE2-1,c
   (pop.sex.age,pop.nat[-1],pop.prov[-1],pop.size[-1],
   pop.sex.age2[-1]),calfun="linear",bounds=c(0.1,10),
   aggregate.index=~HId,compress=F)
tmrb100[i]=proc.time()-ptm;time

ptm=proc.time()
mrb<-as.svrepdesign(svd,type="mrbbootstrap",replicates
   =200,mse=T)
calmrb<-calibrate(mrb,~SEXgAGE+NAT+PROV+SIZE+SEXgAGE2-1,c
   (pop.sex.age,pop.nat[-1],pop.prov[-1],pop.size[-1],
   pop.sex.age2[-1]),calfun="linear",bounds=c(0.1,10),
   aggregate.index=~HId,compress=F)
tmrb200[i]=proc.time()-ptm;print(time)
}

temps<-c(4.83,8.74,8.90,10.53,55.14,26.01)
col6<-c(3,"aquamarine","dodgerblue","dodgerblue4",2,"gold
   ")
bp<-barplot(rev(temps),col=rev(col6),xlim=c(0,60),main="
   Temps␣mitja␣d'execucio",horiz=T)
text(x= rev(temps)/2, y= bp, labels=as.character(round(
   rev(temps),1)), xpd=TRUE,cex=1.15)
legend(38,7.4,c("Taylor","BRR","BRR␣Fay","BRR40","
   Jackknife","Bootstrap"),pch=15,col=col6)
mtext("Metode", 2, line=1.2, cex = 1.2)
mtext("Temps(s)", 1, line=2.4, cex = 1.2)
```

Apèndix F
Codi R - Resultats

1. Taxes

```
#TAXA D'ACTIVITAT:
TaxaAG<-svymean(~ACTIN+EMPUN+JLESS,cal)
TaxaAG<-cbind(data.frame(TaxaAG),confint(TaxaAG),cv(
   TaxaAG)*100)[2,]
TaxaAG[1:4]=TaxaAG[1:4]*100
colnames(TaxaAG)<-c("Taxa(%)","SE(%)", "IC␣2.5%","IC␣
   97.5%","CV(%)");rownames(TaxaAG)<-c("Taxa␣d'activitat
   ")

TaxaAS<-svyby(~ACTIN,~SEX,cal,svymean,na.rm.y=T, vartype=
   c("se","ci","ci","cv"))[,c(3,5,7,9,11)]*100
colnames(TaxaAS)<-c("Taxa(%)","SE(%)", "IC␣2.5%","IC␣
   97.5%","CV(%)");rownames(TaxaAS)<-c("Home","Dona")

TaxaAN<-svyby(~ACTIN,~NAT,cal,svymean,na.rm.y=T, vartype=
   c("se","ci","ci","cv"))[,c(3,5,7,9,11)]*100
colnames(TaxaAN)<-c("Taxa(%)","SE(%)", "IC␣2.5%","IC␣
   97.5%","CV(%)");rownames(TaxaAN)<-c("Espanyola","No␣
   espanyola")

TaxaA4<-svyby(~ACTIN,~AGE4,cal,svymean,na.rm.y=T, vartype
   =c("se","ci","ci","cv"))[,c(3,5,7,9,11)]*100
colnames(TaxaA4)<-c("Taxa(%)","SE(%)", "IC␣2.5%","IC␣
   97.5%","CV(%)");rownames(TaxaA4)<-c("16-19","20-24","
   25-54","55i+")

TaxaAP<-svyby(~ACTIN,~PROV,cal,svymean,na.rm.y=T, vartype
   =c("se","ci","ci","cv"))[,c(3,5,7,9,11)]*100
```

```
colnames(TaxaAP)<-c("Taxa(%)","SE(%)", "IC␣2.5%","IC␣
   97.5%","CV(%)");rownames(TaxaAP)<-c("Barcelona","
   Girona","Lleida","Tarragona")

TaxaAF<-svyby(~ACTIN,~EDUC,cal,svymean,na.rm.y=T, vartype
   =c("se","ci","ci","cv"))[,c(3,5,7,9,11)]*100
colnames(TaxaAF)<-c("Taxa(%)","SE(%)", "IC␣2.5%","IC␣
   97.5%","CV(%)");rownames(TaxaAF)<-c("Analfabets␣i␣
   educacio␣primaria","Educacio␣secundaria␣1a.␣etapa","
   Educacio␣secundaria␣2a.␣etapa","Educacio␣superior")


# OCUPACIo
TaxaOG<-svymean(~EMPOP,cal)
TaxaOG<-cbind(data.frame(TaxaOG),confint(TaxaOG),cv(
   TaxaOG)*100)[1,]
TaxaOG[1:4]=TaxaOG[1:4]*100
colnames(TaxaOG)<-c("Taxa(%)","SE(%)", "IC␣2.5%","IC␣
   97.5%","CV(%)");rownames(TaxaOG)<-c("Taxa␣d'ocupacio"
   )

TaxaOS<-svyby(~EMPOP,~SEX,cal,svymean,na.rm.y=T, vartype=
   c("se","ci","ci","cv"))[,c(2,5,8,11,14)]*100
colnames(TaxaOS)<-c("Taxa(%)","SE(%)", "IC␣2.5%","IC␣
   97.5%","CV(%)");rownames(TaxaOS)<-c("Home","Dona")

TaxaON<-svyby(~EMPOP,~NAT,cal,svymean,na.rm.y=T, vartype=
   c("se","ci","ci","cv"))[,c(2,5,8,11,14)]*100 #No la
   publiquen
colnames(TaxaON)<-c("Taxa(%)","SE(%)", "IC␣2.5%","IC␣
   97.5%","CV(%)");rownames(TaxaON)<-c("Espanyola","No␣
   espanyola")

TaxaO4<-svyby(~EMPOP,~AGE4,cal,svymean,na.rm.y=T, vartype
   =c("se","ci","ci","cv"))[,c(2,5,8,11,14)]*100
colnames(TaxaO4)<-c("Taxa(%)","SE(%)", "IC␣2.5%","IC␣
   97.5%","CV(%)");rownames(TaxaO4)<-c("16-19","20-24","
   25-54","55i+")

TaxaOP<-svyby(~EMPOP,~PROV,cal,svymean,na.rm.y=T, vartype
   =c("se","ci","ci","cv"))[,c(2,5,8,11,14)]*100
colnames(TaxaOP)<-c("Taxa(%)","SE(%)", "IC␣2.5%","IC␣
   97.5%","CV(%)");rownames(TaxaOP)<-c("Barcelona","
   Girona","Lleida","Tarragona")
```

```
TaxaOF<-svyby(~EMPOP,~EDUC,cal,svymean,na.rm.y=T, vartype
   =c("se","ci","ci","cv"))[,c(2,5,8,11,14)]*100
colnames(TaxaOF)<-c("Taxa(%)","SE(%)", "IC␣2.5%","IC␣
   97.5%","CV(%)");rownames(TaxaOF)<-c("Analfabets␣i␣
   educacio␣primaria","Educacio␣secundaria␣1a.␣etapa","
   Educacio␣secundaria␣2a.␣etapa","Educacio␣superior")


#ATUR:

TaxaPG<-svymean(~UNEMP,na.rm=T,cal)
TaxaPG<-cbind(data.frame(TaxaPG),confint(TaxaPG),cv(
   TaxaPG)*100)[2,]
TaxaPG[1:4]=TaxaPG[1:4]*100
colnames(TaxaPG)<-c("Taxa(%)","SE(%)", "IC␣2.5%","IC␣
   97.5%","CV(%)");rownames(TaxaPG)<-c("Taxa␣d'atur")

TaxaPS<-svyby(~UNEMP,na.rm=T,~SEX,cal,svymean,na.rm.y=T,
   vartype=c("se","ci","ci","cv"))[,c(2,5,8,11,14)+1]*
   100
colnames(TaxaPS)<-c("Taxa(%)","SE(%)", "IC␣2.5%","IC␣
   97.5%","CV(%)");rownames(TaxaPS)<-c("Home","Dona")

TaxaPN<-svyby(~UNEMP,na.rm=T,~NAT,cal,svymean,na.rm.y=T,
   vartype=c("se","ci","ci","cv"))[,c(2,5,8,11,14)+1]*
   100
colnames(TaxaPN)<-c("Taxa(%)","SE(%)", "IC␣2.5%","IC␣
   97.5%","CV(%)");rownames(TaxaPN)<-c("Espanyola","No␣
   espanyola")

TaxaP4<-svyby(~UNEMP,na.rm=T,~AGE4,cal,svymean,na.rm.y=T,
    vartype=c("se","ci","ci","cv"))[,c(2,5,8,11,14)+1]*
   100
colnames(TaxaP4)<-c("Taxa(%)","SE(%)", "IC␣2.5%","IC␣
   97.5%","CV(%)");rownames(TaxaP4)<-c("16-19","20-24","
   25-54","55i+")

TaxaPP<-svyby(~UNEMP,na.rm=T,~PROV,cal,svymean,na.rm.y=T,
    vartype=c("se","ci","ci","cv"))[,c(2,5,8,11,14)+1]*
   100
colnames(TaxaPP)<-c("Taxa(%)","SE(%)", "IC␣2.5%","IC␣
   97.5%","CV(%)");rownames(TaxaPP)<-c("Barcelona","
   Girona","Lleida","Tarragona")
```

```
TaxaPF<-svyby(~UNEMP,na.rm=T,~EDUC,cal,svymean,na.rm.y=T,
    vartype=c("se","ci","ci","cv"))[,c(2,5,8,11,14)+1]*
   100
colnames(TaxaPF)<-c("Taxa(%)","SE(%)", "IC 2.5%","IC
   97.5%","CV(%)");rownames(TaxaPF)<-c("Analfabets i
   educacio primaria","Educacio secundaria 1a. etapa","
   Educacio secundaria 2a. etapa","Educacio superior")
#### Poblacio activa, ocupada i desocupada segons el sexe
   , l'edat, el nivell
# de formacio assolit i la nacionalitat, entre altres.
svyby(~EMPOP,~PROV,cal,svytotal)

plot(ecdf(factors$INE))
lines(ecdf(factors$calib),col=2)
```

2. Totals poblacionals

```
##POBLACIo ACTIVA
PobAG<-svytotal(~ACTIN,cal)
PobAG<-cbind(data.frame(PobAG),confint(PobAG),cv(PobAG)*
   100)[2,]
PobAG[1:4]=PobAG[1:4]
colnames(PobAG)<-c("Total","SE", "IC 2.5%","IC 97.5%","CV
   (%)");rownames(PobAG)<-c("Poblacio activa")

PobAS<-svyby(~ACTIN,~SEX,cal,svytotal,na.rm.y=T, vartype=
   c("se","ci","ci","cvpct"))[,c(3,5,7,9,11)]
colnames(PobAS)<-c("Total","SE", "IC 2.5%","IC 97.5%","CV
   (%)");rownames(PobAS)<-c("Home","Dona")

PobAN<-svyby(~ACTIN,~NAT,cal,svytotal,na.rm.y=T, vartype=
   c("se","ci","ci","cvpct"))[,c(3,5,7,9,11)]
colnames(PobAN)<-c("Total","SE", "IC 2.5%","IC 97.5%","CV
   (%)");rownames(PobAN)<-c("Espanyola","No espanyola")

PobA4<-svyby(~ACTIN,~AGE4,cal,svytotal,na.rm.y=T, vartype
   =c("se","ci","ci","cvpct"))[,c(3,5,7,9,11)]
colnames(PobA4)<-c("Total","SE", "IC 2.5%","IC 97.5%","CV
   (%)");rownames(PobA4)<-c("16-19","20-24","25-54","55i
   +")

PobAP<-svyby(~ACTIN,~PROV,cal,svytotal,na.rm.y=T, vartype
   =c("se","ci","ci","cvpct"))[,c(3,5,7,9,11)]
```

```
colnames(PobAP)<-c("Total","SE", "IC␣2.5%","IC␣97.5%","CV
   (%)");rownames(PobAP)<-c("Barcelona","Girona","Lleida
   ","Tarragona")

PobAF<-svyby(~ACTIN,~EDUC,cal,svytotal,na.rm.y=T, vartype
   =c("se","ci","ci","cvpct"))[,c(3,5,7,9,11)]
colnames(PobAF)<-c("Total","SE", "IC␣2.5%","IC␣97.5%","CV
   (%)");rownames(PobAF)<-c("Analfabets␣i␣educacio␣
   primaria","Educacio␣secundaria␣1a.␣etapa","Educacio␣
   secundaria␣2a.␣etapa","Educacio␣superior")


PobAE<-svyby(~ACTIN,~ECO,cal,svytotal,na.rm.y=T, vartype=
   c("se","ci","ci","cvpct"))[,c(3,5,7,9,11)]
colnames(PobAE)<-c("Total","SE", "IC␣2.5%","IC␣97.5%","CV
   (%)");rownames(PobAE)<-c("Agricultura","Industria","
   Construccio","Serveis")

PobAB<-svyby(~ACTIN,~BRAN,cal,svytotal,na.rm.y=T, vartype
   =c("se","ci","ci","cvpct"))[,c(3,5,7,9,11)]
colnames(PobAB)<-c("Total","SE", "IC␣2.5%","IC␣97.5%","CV
   (%)")

# POBLACIo OCUPADA
PobOG<-svytotal(~EMPOP,cal)
PobOG<-cbind(data.frame(PobOG),confint(PobOG),cv(PobOG)*
   100)[1,]
PobOG[1:4]=PobOG[1:4]
colnames(PobOG)<-c("Total","SE", "IC␣2.5%","IC␣97.5%","CV
   (%)");rownames(PobOG)<-c("Poblacio␣ocupada")

PobOS<-svyby(~EMPOP,~SEX,cal,svytotal,na.rm.y=T, vartype=
   c("se","ci","ci","cvpct"))[,c(2,5,8,11,14)]
colnames(PobOS)<-c("Total","SE", "IC␣2.5%","IC␣97.5%","CV
   (%)");rownames(PobOS)<-c("Home","Dona")

PobON<-svyby(~EMPOP,~NAT,cal,svytotal,na.rm.y=T, vartype=
   c("se","ci","ci","cvpct"))[,c(2,5,8,11,14)] #No la
   publiquen
colnames(PobON)<-c("Total","SE", "IC␣2.5%","IC␣97.5%","CV
   (%)");rownames(PobON)<-c("Espanyola","No␣espanyola")

PobO4<-svyby(~EMPOP,~AGE4,cal,svytotal,na.rm.y=T, vartype
   =c("se","ci","ci","cvpct"))[,c(2,5,8,11,14)]
```

```
colnames(PobO4)<-c("Total","SE", "IC␣2.5%","IC␣97.5%","CV
   (%)");rownames(PobO4)<-c("16-19","20-24","25-54","55i
   +")

PobOP<-svyby(~EMPOP,~PROV,cal,svytotal,na.rm.y=T, vartype
   =c("se","ci","ci","cvpct"))[,c(2,5,8,11,14)]
colnames(PobOP)<-c("Total","SE", "IC␣2.5%","IC␣97.5%","CV
   (%)");rownames(PobOP)<-c("Barcelona","Girona","Lleida
   ","Tarragona")

PobOF<-svyby(~EMPOP,~EDUC,cal,svytotal,na.rm.y=T, vartype
   =c("se","ci","ci","cvpct"))[,c(2,5,8,11,14)]
colnames(PobOF)<-c("Total","SE", "IC␣2.5%","IC␣97.5%","CV
   (%)");rownames(PobOF)<-c("Analfabets␣i␣educacio␣
   primaria","Educacio␣secundaria␣1a.␣etapa","Educacio␣
   secundaria␣2a.␣etapa","Educacio␣superior")

PobOE<-svyby(~EMPOP,~ECO,cal,svytotal,na.rm.y=T, vartype=
   c("se","ci","ci","cvpct"))[,c(2,5,8,11,14)]
colnames(PobOE)<-c("Total","SE", "IC␣2.5%","IC␣97.5%","CV
   (%)");rownames(PobOE)<-c("Agricultura","Industria","
   Construccio","Serveis")

PobOB<-svyby(~EMPOP,~BRAN,cal,svytotal,na.rm.y=T, vartype
   =c("se","ci","ci","cvpct"))[,c(2,5,8,11,14)]
colnames(PobOB)<-c("Total","SE", "IC␣2.5%","IC␣97.5%","CV
   (%)");


#POBLACIo DESOCUPADA
PobPG<-svytotal(~UNEMP,na.rm=T,cal)
PobPG<-cbind(data.frame(PobPG),confint(PobPG),cv(PobPG)*
   100)[2,]
PobPG[1:4]=PobPG[1:4]
colnames(PobPG)<-c("Total","SE", "IC␣2.5%","IC␣97.5%","CV
   (%)");rownames(PobPG)<-c("Poblacio␣desocupada")

PobPS<-svyby(~UNEMP,na.rm=T,~SEX,cal,svytotal,na.rm.y=T,
   vartype=c("se","ci","ci","cvpct"))[,c(2,5,8,11,14)+1]
colnames(PobPS)<-c("Total","SE", "IC␣2.5%","IC␣97.5%","CV
   (%)");rownames(PobPS)<-c("Home","Dona")

PobPN<-svyby(~UNEMP,na.rm=T,~NAT,cal,svytotal,na.rm.y=T,
   vartype=c("se","ci","ci","cvpct"))[,c(2,5,8,11,14)+1]
```

```
colnames(PobPN)<-c("Total","SE", "IC␣2.5%","IC␣97.5%","CV
   (%)");rownames(PobPN)<-c("Espanyola","No␣espanyola")

PobP4<-svyby(~UNEMP,na.rm=T,~AGE4,cal,svytotal,na.rm.y=T,
    vartype=c("se","ci","ci","cvpct"))[,c(2,5,8,11,14)
   +1]
colnames(PobP4)<-c("Total","SE", "IC␣2.5%","IC␣97.5%","CV
   (%)");rownames(PobP4)<-c("16-19","20-24","25-54","55i
   +")

PobPP<-svyby(~UNEMP,na.rm=T,~PROV,cal,svytotal,na.rm.y=T,
    vartype=c("se","ci","ci","cvpct"))[,c(2,5,8,11,14)
   +1]
colnames(PobPP)<-c("Total","SE", "IC␣2.5%","IC␣97.5%","CV
   (%)");rownames(PobPP)<-c("Barcelona","Girona","Lleida
   ","Tarragona")

PobPF<-svyby(~UNEMP,na.rm=T,~EDUC,cal,svytotal,na.rm.y=T,
    vartype=c("se","ci","ci","cvpct"))[,c(2,5,8,11,14)
   +1]
colnames(PobPF)<-c("Total","SE", "IC␣2.5%","IC␣97.5%","CV
   (%)");rownames(PobPF)<-c("Analfabets␣i␣educacio␣
   primaria","Educacio␣secundaria␣1a.␣etapa","Educacio␣
   secundaria␣2a.␣etapa","Educacio␣superior")

PobPE<-svyby(~UNEMP,na.rm=T,~ECO,cal,svytotal,na.rm.y=T,
   vartype=c("se","ci","ci","cvpct"))[,c(2,5,8,11,14)+1]
colnames(PobPE)<-c("Total","SE", "IC␣2.5%","IC␣97.5%","CV
   (%)");rownames(PobPE)<-c("Agricultura","Industria","
   Construccio","Serveis")
```